

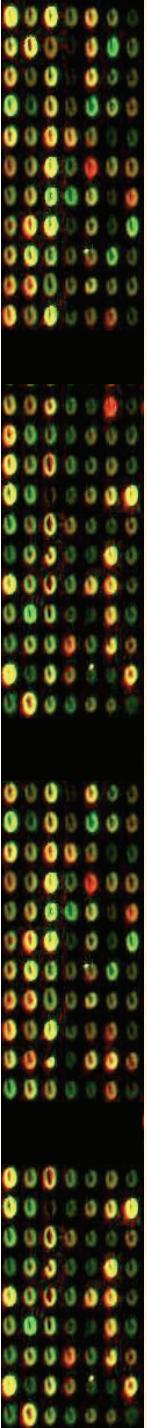
From Community Structure to Functions: GeoChip Development and Its Applications to Bioremediation

Jizhong (Joe) Zhou

jzhou@ou.edu; 405-325-6073

**Institute for Environmental Genomics, Department of
Botany and Microbiology, University of Oklahoma,
Norman, OK 73019**

**DOE ERSP Annual PI Meeting, Lansdowne, VA, April 7-
9, 2008**



Outline

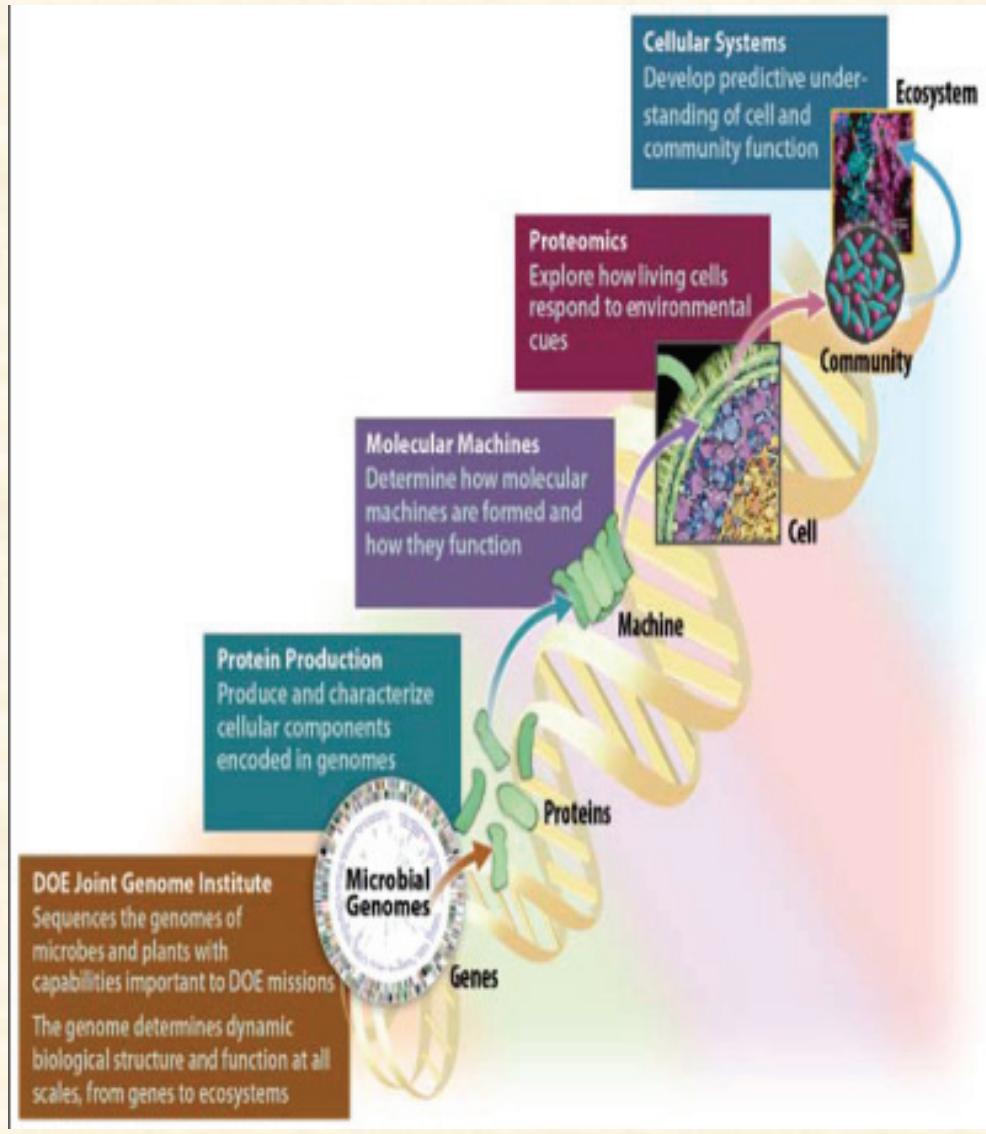
- Brief description of GeoChip technologies
- GeoChip 3.0
 - Functional Gene categories
 - Automatic pipeline
- Applications of GeoChip 2.0
 - Responses of ethanol injection experiments
 - Old Rifle
- Challenges and future perspectives



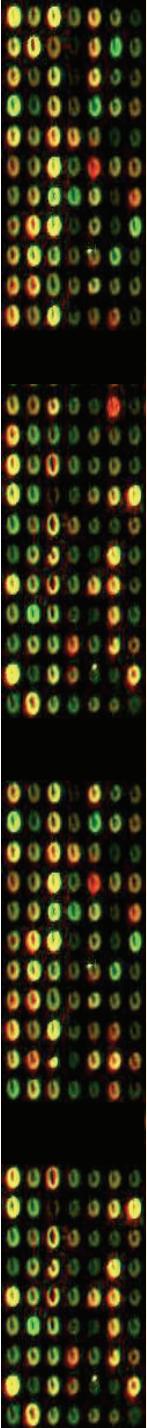
Objectives

- Program mission
 - Understand the hydrogeological and biogeochemical factors that govern the distribution and functioning of subsurface microbial communities
 - Develop techniques to quantitatively identify and quantitate active members of subsurface microbial communities and relate growth and activity to rates of biogeochemical reactions
- Goal of this study
 - Develop and use microarray technologies to provide a fundamental understanding of the structure, functions, activities of microbial communities at DOE sites

Some Grand Challenges in 21st Century Biology

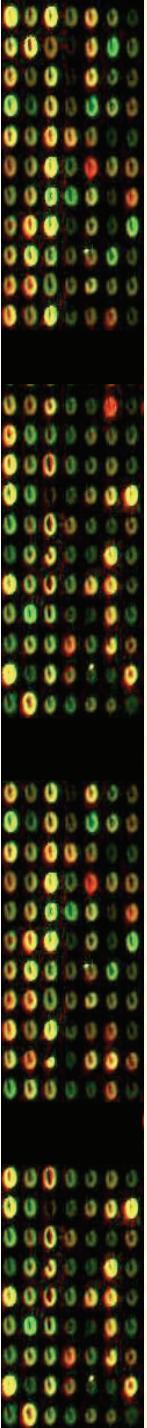


- **Linking genomics to ecology**
 - Linking genomics to ecological processes and functions
 - Responses to CO₂, global warming and water precipitation
- **Linking biodiversity to ecosystem functions**
- **Informational scaling**
 - From cells to individuals, populations, communities, ecosystems and biosphere.
 - Spatial, temporal



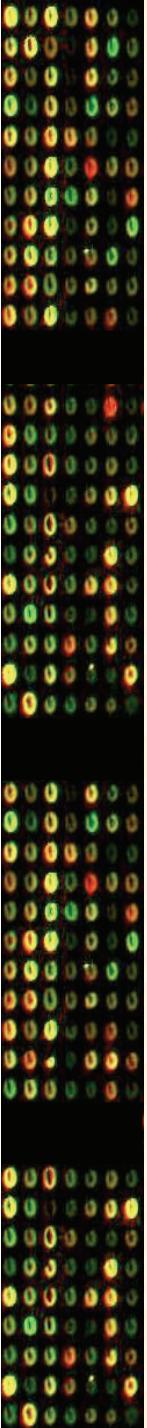
Challenges in Microbial Ecology

- **Extremely high diversity:** 5000 species/g soil
- **Uncultured:** 99% of the microbial species are uncultured
- **Small:** Difficult to see
- **Limited use of morphology:** It could not provide enough information and resolution to differentiate different microorganisms
- **Molecular biology:** Greatly advance our understanding of microbial diversity
- **Limitations of conventional molecular biology tools:** Slow, complicated, labor-intensive, insensitive, unquantitative, and/or small-scale.
- **Genomics technologies:** Revolution in microbial ecology



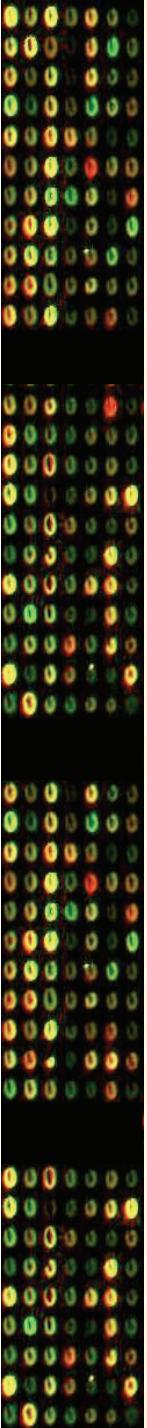
High throughput approaches

- Open format detection
 - High throughput Sequencing.;
 - 454 sequencing, 250bp, 60-100mb/run
 - Solexa, SOLiD: 35bp, 1-2 gb/run
 - Proteomics
- Close format
 - PhyloChip: 16S genes
 - GeoChip: functional genes



Comparisons between open format and close format detection

	Open format	Close format
Affected by random sampling	Yes	No
Effects by dominant populations	Yes	No
Finding new things	Yes	No
Reliable comparison across samples	?	Yes

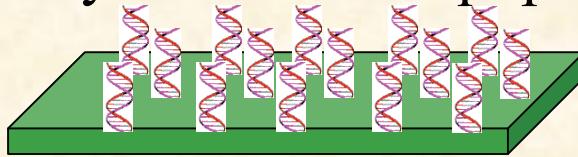


Main advantages of GeoChip compared to other approaches

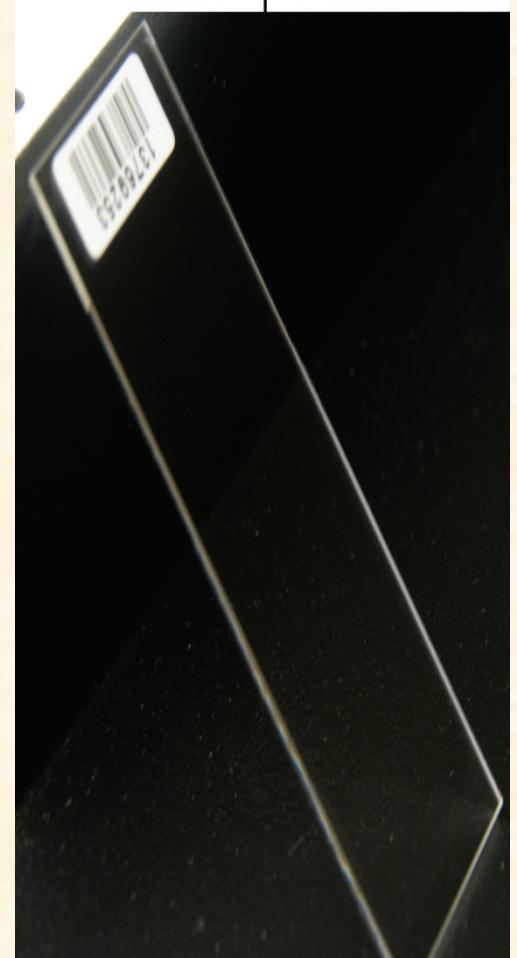
- **Detecting functions:** Geochemical processes
- **Higher resolution:** Species-strain level resolution
- **Quantitative:** no PCR is involved

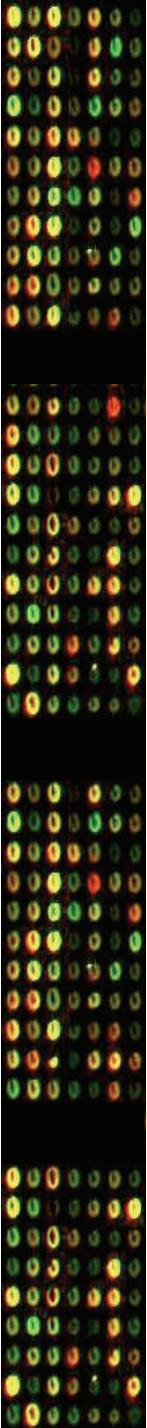
GeoChip or Functional Gene Arrays (FGAs)

- **Microarrays:** Glass slides or other solid surface containing thousands of genes arrayed by automated equipment.



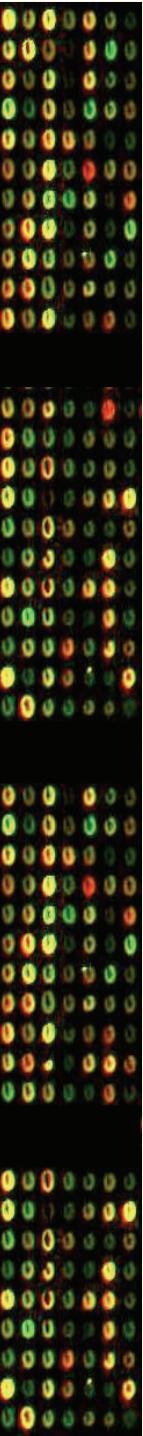
- FGAs contain probes from the genes involved in various geochemical, ecological and environmental processes.
 - C, N, S, P cyclings
 - Organic contaminant degradation
 - Metal resistance and reduction
- Typical format: 50mer oligonucleotide arrays
- Useful for studying microbial communities
 - Functional gene diversity and activity
 - Limited phylogenetic diversity.





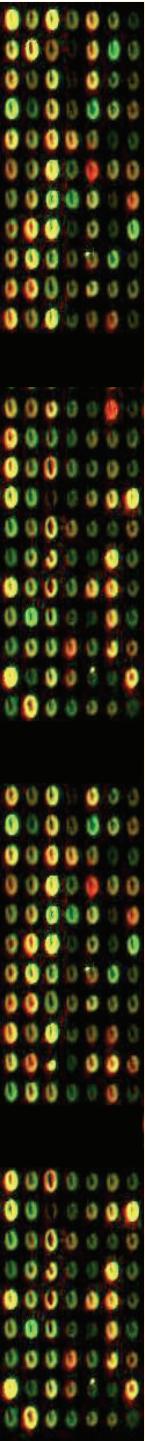
Pioneering advances in microarray-based technologies to address challenges in microbial community genomics

- Challenges:
 - Specificity: Environmental sequence divergences.
 - Sensitivity: Low biomass.
 - Quantification:
 - Existence of contaminants: Humic materials, organic contaminants, metals and radionuclides.
- Solutions
 - Developing different types of microarrays and novel chemistry to address different levels of specificity.
 - Developing novel signal amplification strategy to increase sensitivity
 - Optimizing microarray protocols for reliable quantification.

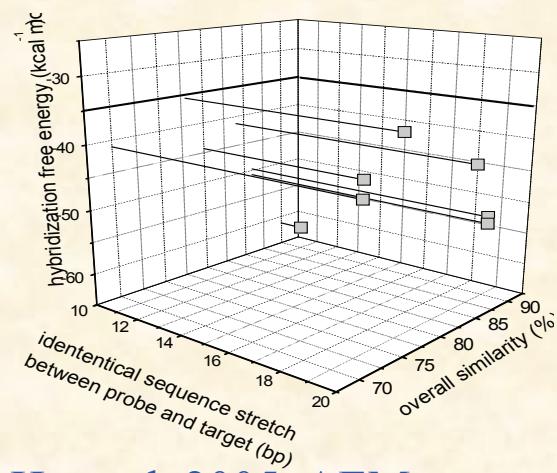
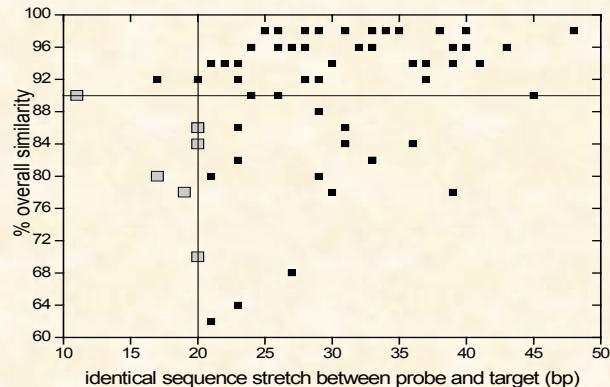


Issues related to specificity, sensitivity and quantitation

- Specificity, sensitivity, quantitation
 - Wu et al. 2001; AEM:67: 5780-5790
 - Rhee et al. 2004, AEM 70:4303-4317
 - Tiquia et al. 2004. BioTechniques 36, 664-675
 - Wu et al. 2004; EST, 38: 6775-6782
 - He et al, 2007; The ISME J, 1: 67-77
 - He and Zhou, 2008, AEM, in press
- Probe design criteria
 - He et al. 2005. AEM. 71:3753-3760
 - Liebich et al. AEM, 72:1688-1691
- New probe designing software: CommOligo
 - Li et al. 2005. Nucl. Acids Res. 33:6114-6123
- Whole community genome amplification (WCGA)
 - Wu et al. 2006. AEM: 72:4931-4941.
- Whole community RNA amplification (WCRA)
 - Gao et al, 2007, AEM: 73: 563-571.
- Review:
 - Gentry et al. 2006, Microbial Ecology, 52: 159-175.
 - Zhou and Thompson, 2002, Curr Opion Biotech: 13:204-207
 - Zhou, 2003; Curr Opion. Microbiol, 6:288-294
- Applications
 - He et al, 2007; The ISME J, 1: 67-77
 - Leigh et al, 2007, The ISME J, 1: 163-179
 - Yergeau et al, 2007, The ISME J, 1: 134-148.
 - Zhou et al. 2008. PNAS, in press



Empirical criteria for designing specific probes

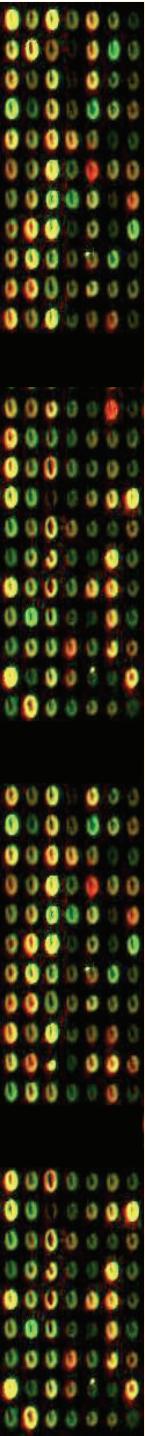


He et al. 2005. AEM.
71:3753-3760
Liebich et al. AEM,
72:1688-1691

- Hybridization condition: 50 C, 50% formamide
- Evaluation of 516 probes with homology between less than 86% and 100% or a common identical sequence stretch of at least 16 bases.
- SNR > 2 as positive

Criteria for designing specific probes

- Similarity: $\leq 90\%$
- Stretches: ≤ 20 bases
- Free energy: ≥ -35 kcal/mol
- Specific hybridization was also observed for some probes with less than 98% similarity to the target sequences, suggesting that strain level of resolution could be possibly achieved with some 50mer probes under the hybridization conditions examined.

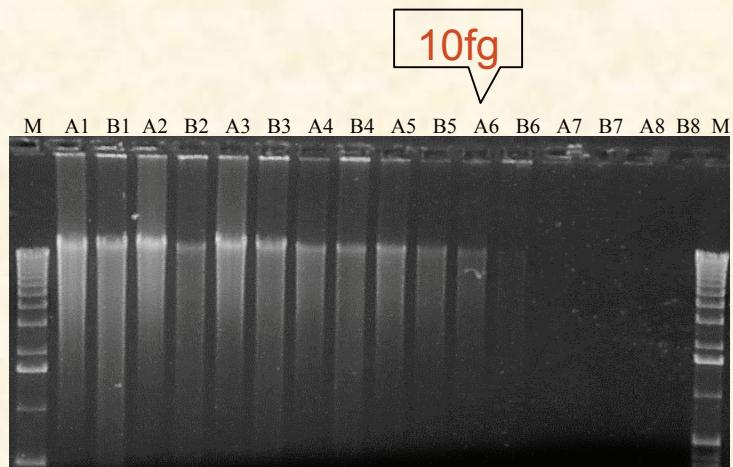


Resolution of 50mer oligo arrays

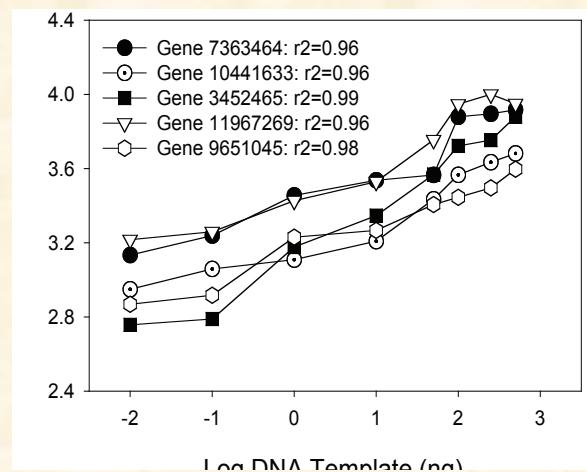
Phylogenetic hierarchies	Sequence similarities (%) \pm standard deviation†					
	<i>DsrAB</i>	<i>nirS</i>	<i>nirK</i>	<i>nifH</i>	<i>amoA</i>	<i>pmoA</i>
Strain	0.93 \pm 0.03	0.93 \pm 0.04	0.91 \pm 0.05	0.93 \pm 0.06	0.99 \pm 0.01	0.95 \pm 0.03
Species	0.73 \pm 0.13	0.70 \pm 0.18	0.76 \pm 0.00	0.82 \pm 0.06	0.75 \pm 0.11	0.79 \pm 0.55
Genus	0.70 \pm 0.09	0.67 \pm 0.16	0.71 \pm 0.07	0.70 \pm 0.07	0.71 \pm 0.07	0.75 \pm 0.15
Family or higher	0.66 \pm 0.13	0.57 \pm 0.14	0.62 \pm 0.14	0.66 \pm 0.10	0.65 \pm 0.13	-

- Based on pure cultures
- FGA can achieve species level identification

Whole community genome amplification (WCGA) approach for increasing hybridization sensitivity



As low as 10fg (2 cells)
can be detected

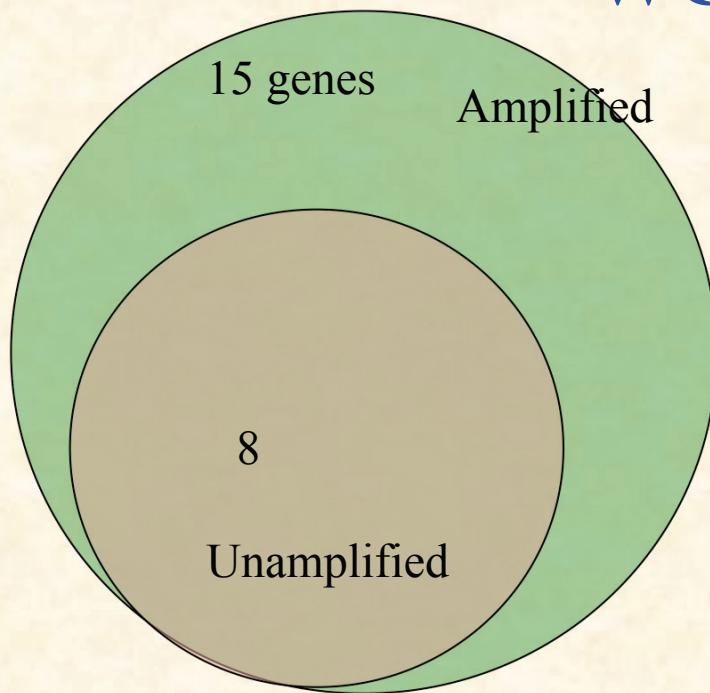


Quantitative after amplification

1-100 ng DNAs are typically used for analysis

Wu et al. 2006. AEM: 72:4931-4941, top 11th most requested paper in AEM.

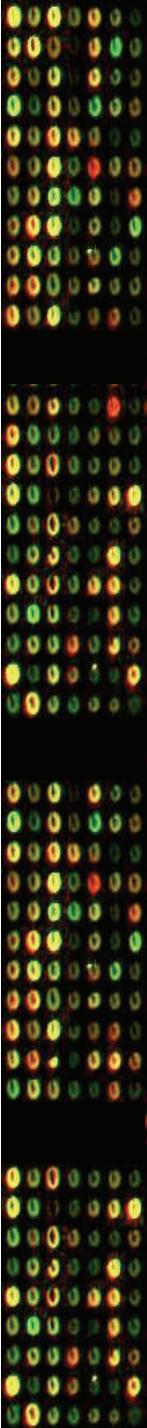
Whole Community RNA Amplification WCRA)



- Fluidized Bioreactor Reactors for removing nitrate.

Gao et al, 2007, AEM: 73: 563-571. Top 8th most requested paper in AEM.

- 5 ug RNA --- umamplified.
- 5 ug amplified RNA from 100 ng RNA as template for amplification
- 8 genes are detected in both amplified and unamplified RNAs
- 7 more genes are detected by RNA amplification



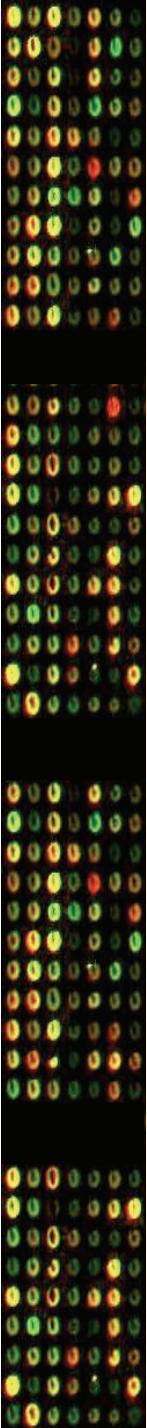
CommOligo --- New oligo probe design program for community analysis

Number and specificity of designed probes (50-mer) by different programs

Programs used	Group sequences of <i>nirS</i> and <i>nirK</i> (842 gene sequences)					
	Total ORFs	ORFs rejected	Probes designed	Specific probe	Non-specific	Group -specific
ArrayOligoSe ector	842	0	842	117	725	0
OligoArray	842	35	807	70	737	0
OligoArray 2.0	842	51	791	35	756	0
OligoPicker	842	657	185	141	44	0
CommOligo	842	512	330	147	0	183

- Useful for both whole genome microarrays and community arrays
- Able to design group-specific probes
- Better performance than other programs

Li et al. 2005. Nucl. Acids Res. 33:6114-6123

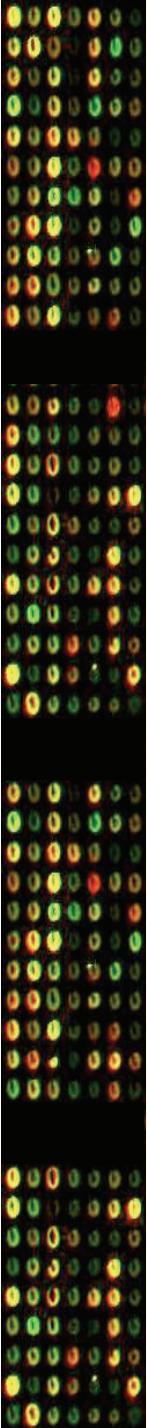


Probes Designed for a Second Generation FGA (GeoChip 2.0)

- Nitrogen cycling: 5089
- Carbon cycling: 9198
- Sulfate reduction: 1006
- Phosphorus utilization: 438
- Organic contaminant degradation: 5359
- Metal resistance and oxidation: 2303
- 300 probes from community sequences

Total: 23,408 genes

- 23,000 probes designed
- Will be very useful for community and ecological studies

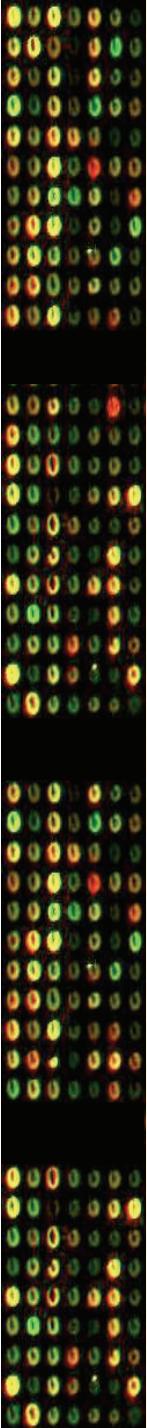


GeoChip for microbial community analysis

He, Z, TJ Gentry, CW Schadt, L Wu, J Liebich, SC Chong, Z Huang, W Wu, B Gu, P Jardine, C Criddle, and J. Zhou. 2007. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. The ISME J. 1: 67-77.

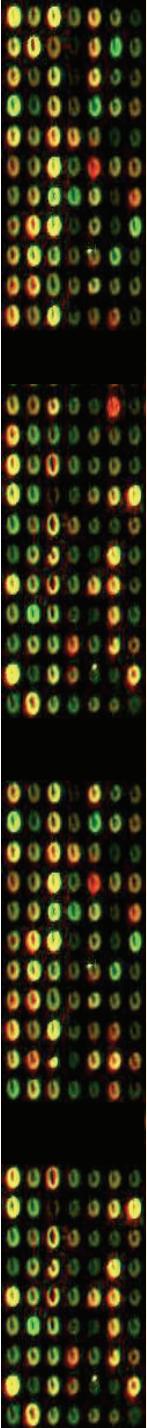
Highlighted by:

- A press release by Nature Press Office
- Reported by many Newspapers
- National Ecology Observatory Networks (NEON), Roadmap
- National Academy of Sciences, Metagenomics report



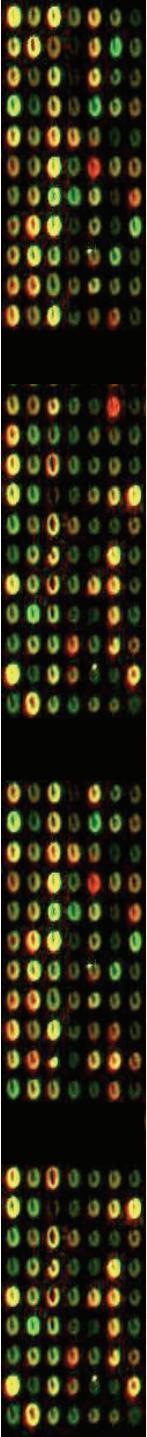
New features for GeoChip 3.0

- GeoChip 3.0 is more comprehensive and more representative. It covers >37,700 gene sequences of 290 gene families.
- Automatically retrieved sequences by key words is verified by HUMMER and then unrelated sequences are removed.
- A software package for sequence retrieval, probe and array design, probe verification, array construction, array data analysis, and information storage.
- Automatic update greatly facilitates the management of such a complicated functional gene array.
- New analysis pipeline with many statistical and mathematical tools



Summary of GeoChip 3.0 probe and sequence information by category

Gene category	No. of gene categories	No. of downloaded sequences	No. of sequences for probe design	Total no. of probes designed	Total no. of CDS covered
Carbon degradation	24	18337	4092	1924	3192
Carbon fixation	5	4682	2218	887	1614
Methane reduction and oxidation	3	4134	1853	447	752
Metal resistance and reduction	43	28820	9625	3510	7021
Nitrogen cycling	12	20800	19229	4006	7334
Organic remediation	197	55598	18650	7093	12843
Phosphorus utilization	2	1876	1441	471	1069
Sulfur cycling	3	2523	2291	1464	1800
Others (e.g. <i>gyrB</i>)	1	8163	5252	1040	2089
Total	290	144,933	64,651	20,842*	37,714



Carbon degradation

Gene/category	Unique probe	Group probe	Total probe	Total covered CDS
Carbon degradation				
acetylglucosaminidase	32	75	107	214
<i>amyA</i>	61	170	231	467
<i>amyX</i>	0	5	5	12
<i>apu</i>	4	2	6	8
<i>ara</i>	21	65	86	174
<i>ara_fungi</i>	23	10	33	50
<i>cda</i>	11	6	17	25
cellobiase	36	41	77	145
endochitinase	199	168	367	606
endoglucanase	64	24	88	109
exochitinase	15	16	31	63
exoglucanase	54	9	63	83
glucoamylase	23	35	58	111
<i>glx</i>	17	4	21	33
isopullulanase	0	1	1	2
<i>lip</i>	25	4	29	39
mannanase	20	9	29	45
<i>mnp</i>	17	2	19	22
<i>nplT</i>	4	16	20	39
pectinase	27	2	29	33
phenol_oxidase	126	81	207	272
<i>pulA</i>	21	88	109	231
<i>xylA</i>	18	72	90	188
xylanase	60	67	127	221
Subtotal	878	972	1850	3192

Carbon fixation and methane metabolism

Gene/category	Unique probe	Group probe	Total probe	Total covered CDS
<u>Carbon fixation</u>				
<i>aclB</i>	20	13	33	53
CODH	13	63	76	138
FTHFS	68	126	194	323
<i>pcc</i>	8	249	257	585
<i>rubisco</i>	139	146	285	515
Subtotal	248	597	845	1614
<u>Methane metabolism</u>				
<i>mcrA</i>	104	106	210	392
<i>mmoX</i>	22	22	44	90
<i>pmoA</i>	85	39	124	270
Subtotal	211	167	378	752

Nitrogen cycling

Gene/category	Unique probe	Group probe	Total probe	Total covered CDS
<u>Nitrogen cycling</u>				
<i>amoA</i>	100	95	195	528
<i>gdh</i>	26	19	45	94
<i>hao</i>	2	4	6	18
<i>napA</i>	11	22	33	83
<i>narG</i>	289	160	449	656
<i>nasA</i>	67	86	153	259
<i>nifH</i>	885	333	1218	2467
<i>nirK</i>	255	143	398	1005
<i>nirS</i>	351	155	506	923
<i>norB</i>	55	25	80	102
<i>nosZ</i>	191	119	310	596
<i>ureC</i>	57	218	275	603
Subtotal	2289	1379	3668	7334

Phosphorus utilization and sulphur cycling

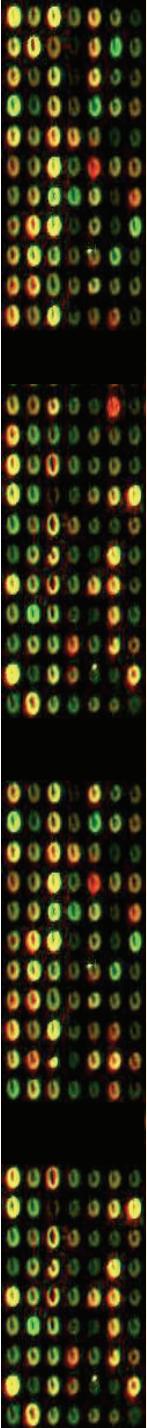
Gene/category	Unique probe	Group probe	Total probe	Total covered CDS
<u>Phosphorus</u>				
<i>ppk</i>	47	67	114	237
<i>ppx</i>	44	296	340	832
Subtotal	91	363	454	1069
<u>Sulphur</u>				
<i>dsrA</i>	595	155	750	954
<i>dsrB</i>	371	131	502	685
<i>sox</i>	47	52	99	161
Subtotal	1013	338	1351	1800

Metal reduction and resistance

Gene/category	Total probe	Total covered CDS
<u>Metal reduction and resistance</u>		
Arsenic resistance	396	803
Cadmium resistance	1254	2808
Chromium resistance	543	1292
Mercury resistance/reduction	292	594
Nickel resistance	42	88
Zinc resistance	1044	2197
Other metals and metalloids	1803	4135
Other metal reduction	413	449
Subtotal	5,787	12,366

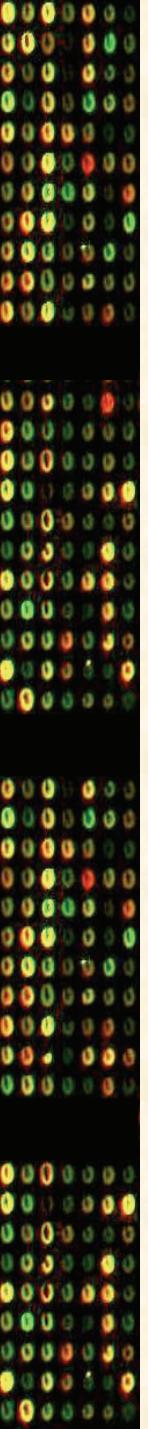
Organic contaminant degradation

Gene/category	Total probe	Total covered CDS
<u>Contaminant degradation</u>		
BTEX and related aromatics	423	3084
Chloronated aromatics	250	473
Nitroaromatics	122	489
Heterocyclic aromatics	38	66
Hydrocarbons (e.g., PAHs)	179	2089
Chloronated solvents	180	360
Pesticides	1258	3083
Other chemicals and by-products	3936	7855
Subtotal	6,386	17,499



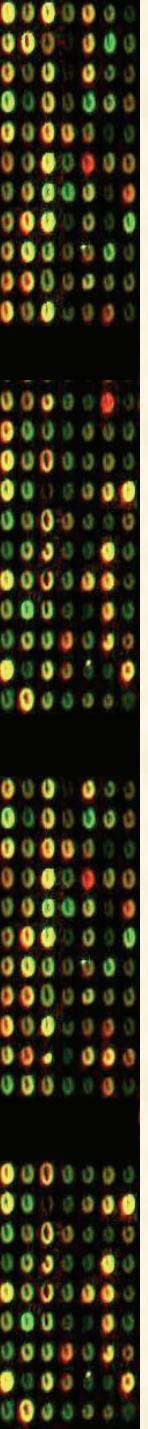
Energy-related metabolism processes

Gene/category	Total probe	Total covered CDS
<u>Energy-related metabolism processes</u>		
Cytochromes	365	365
Hydrogenase	48	85
Subtotal	413	450



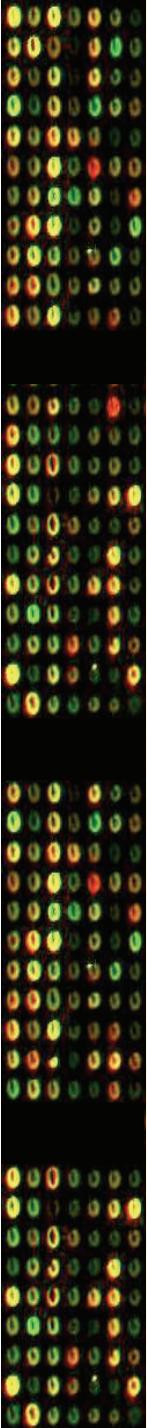
Two major types of applications

- Addressing site-specific research questions:
 - Need understanding community diversity first.
- Profiling microbial communities as a generic tool



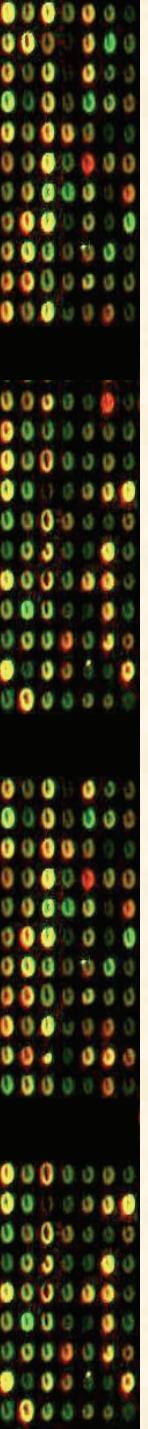
Examples of most recent applications

- Groundwaters
 - Monitoring bioremediation processes: Ur, Cr
 - Impacts of contaminants on microbial communities
- Soils
 - Grass land soils: effects of plant diversity and climate changes on soil communities
 - Agricultural soils: tillage, no tillage
 - Oil-contaminated soils
- Aquatic environments
 - Hydrothermal vents
 - Marine sediments
 - River sediments
- Bioreactors
 - Wastewater treatments
 - Biohydrogen
 - Microbial fuel cell
- Bioleaching



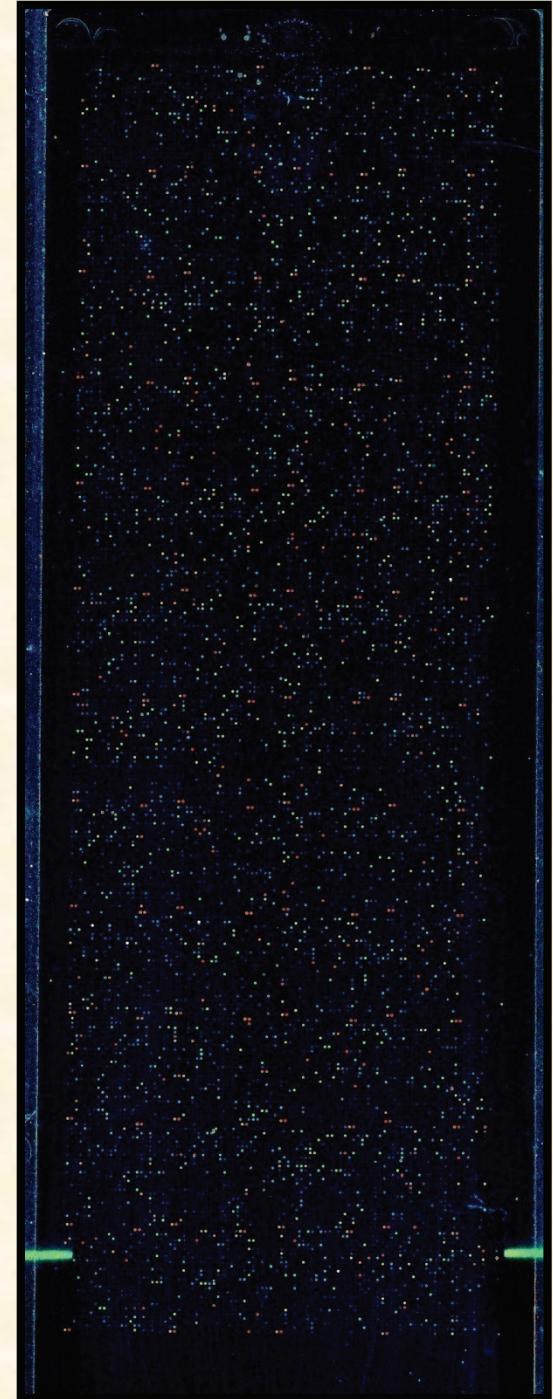
Applications to Bioremediation

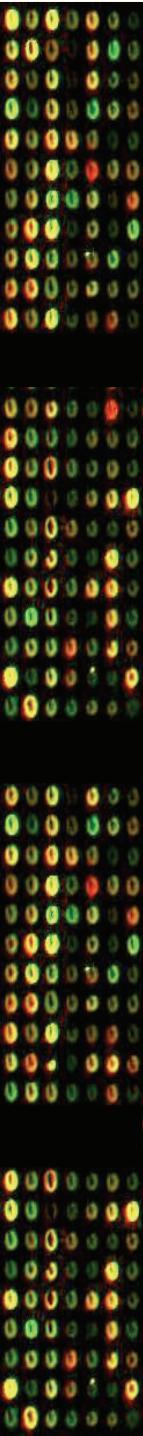
- Oak Ridge
 - Impacts of contaminants on groundwater microbial community, **natural attenuation**
 - **Responses of microbial communities to ethanol treatments**
 - Microbial community composition in soil microcosms during U-reduction, U-oxidation, and mobilization phases
- Hanford
 - Impacts of contaminants on microbial communities, 5 samples from J. Fredickson
 - Responses of microbial communities to the addition of Hydrogen Release Compound (HRC)
- Old Rifle
 - In-situ U(VI) bioremediation under sulfate-reducing and Fe-reducing conditions
- Lake Depue
 - Metal contamination



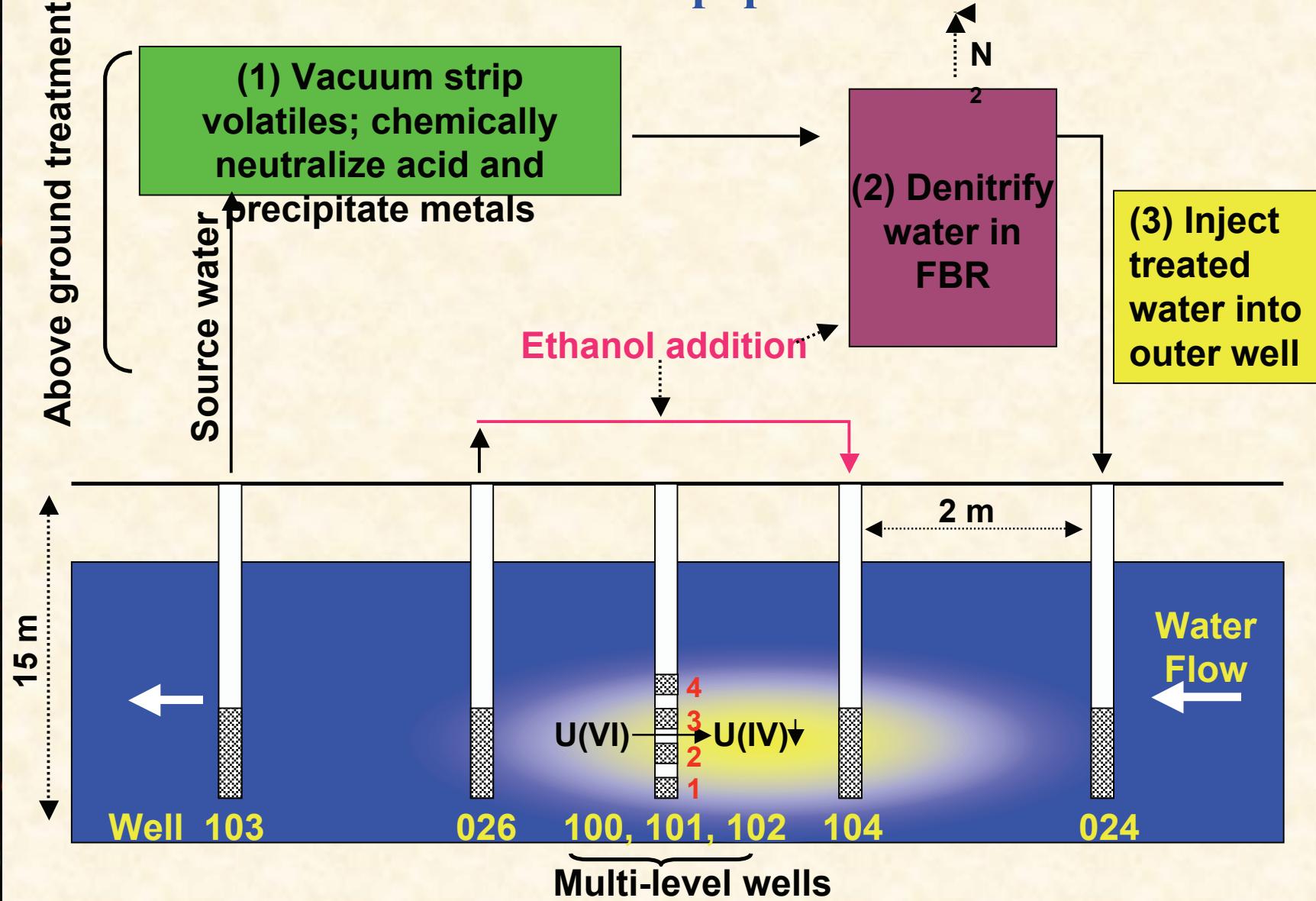
Overview of Microarray Analysis

- DNA extraction from environmental samples, multiple samples, times
- Whole Community Rolling Circle Amplification (**1-100ng DNA**)
- Label DNA with Cy5
- Hybridization to GeoChip at 42, 45 or 50C with 50% formamide
- Data processing with automatic pipeline
- Statistical analysis
- Data interpretation



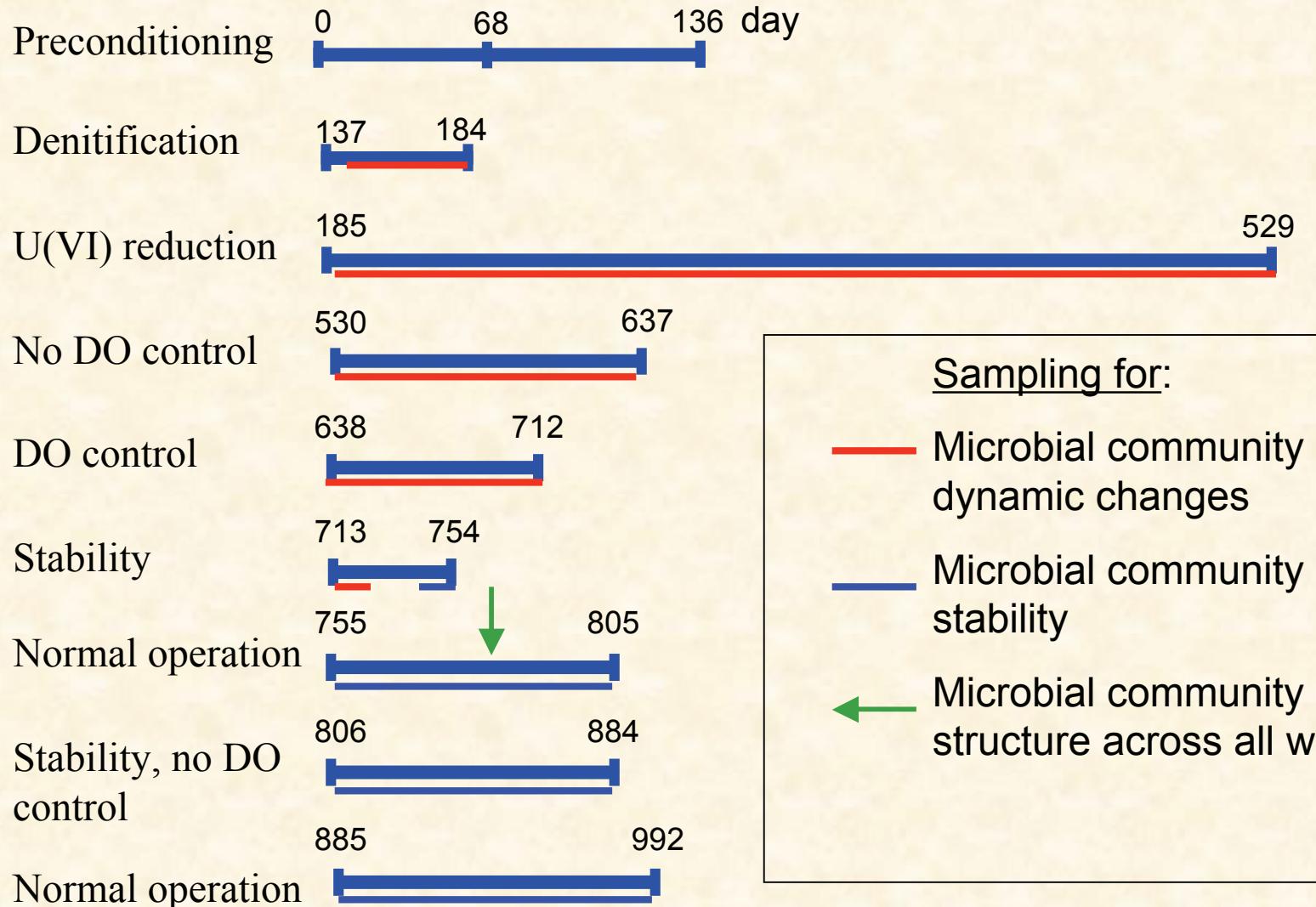


Biostimulation of microbial populations for Ur removal

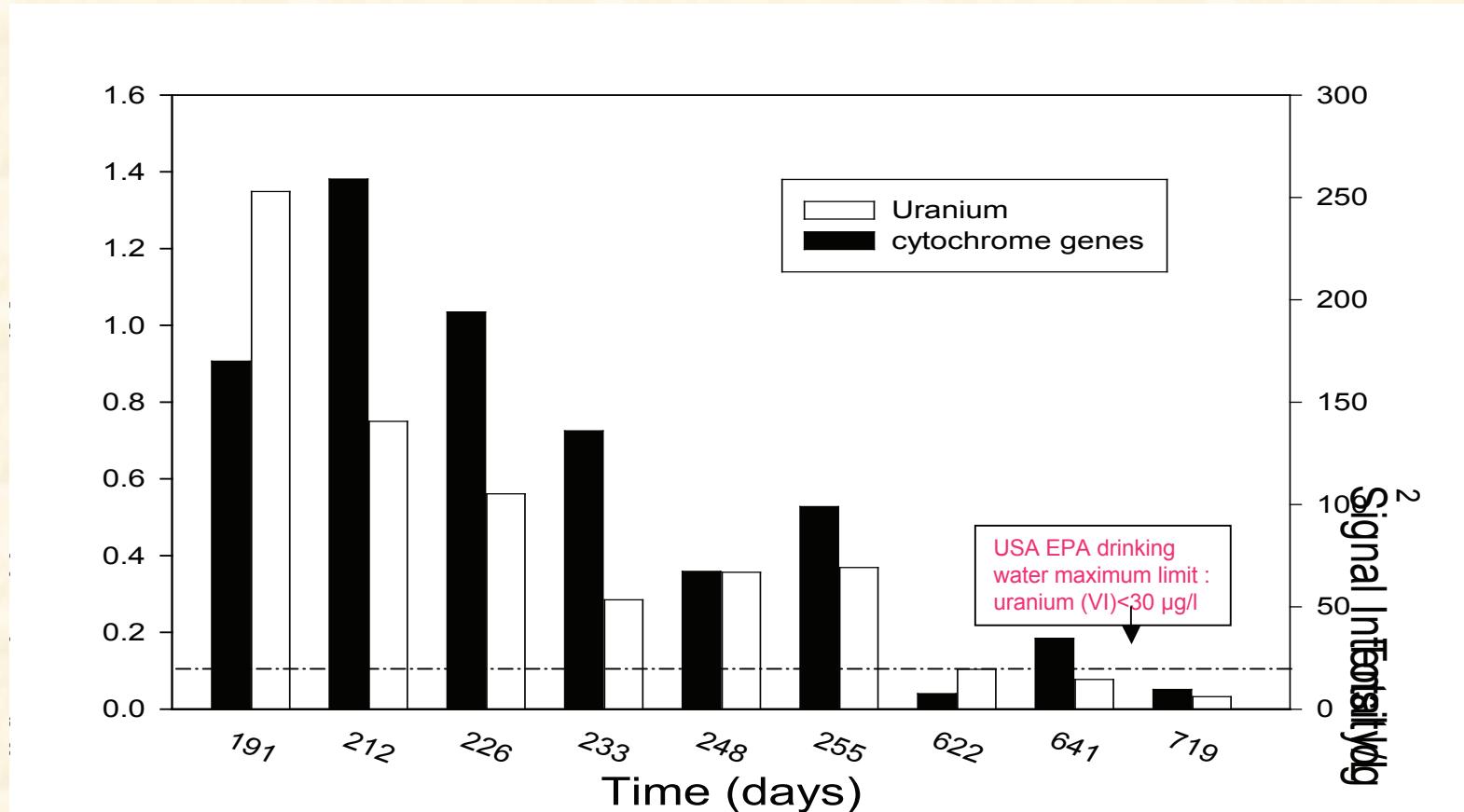


- Above ground denitrification and neutralization of groundwater
- *in situ* biostimulation with ethanol and reduction of U(VI)

Operational Periods During Bioremediation

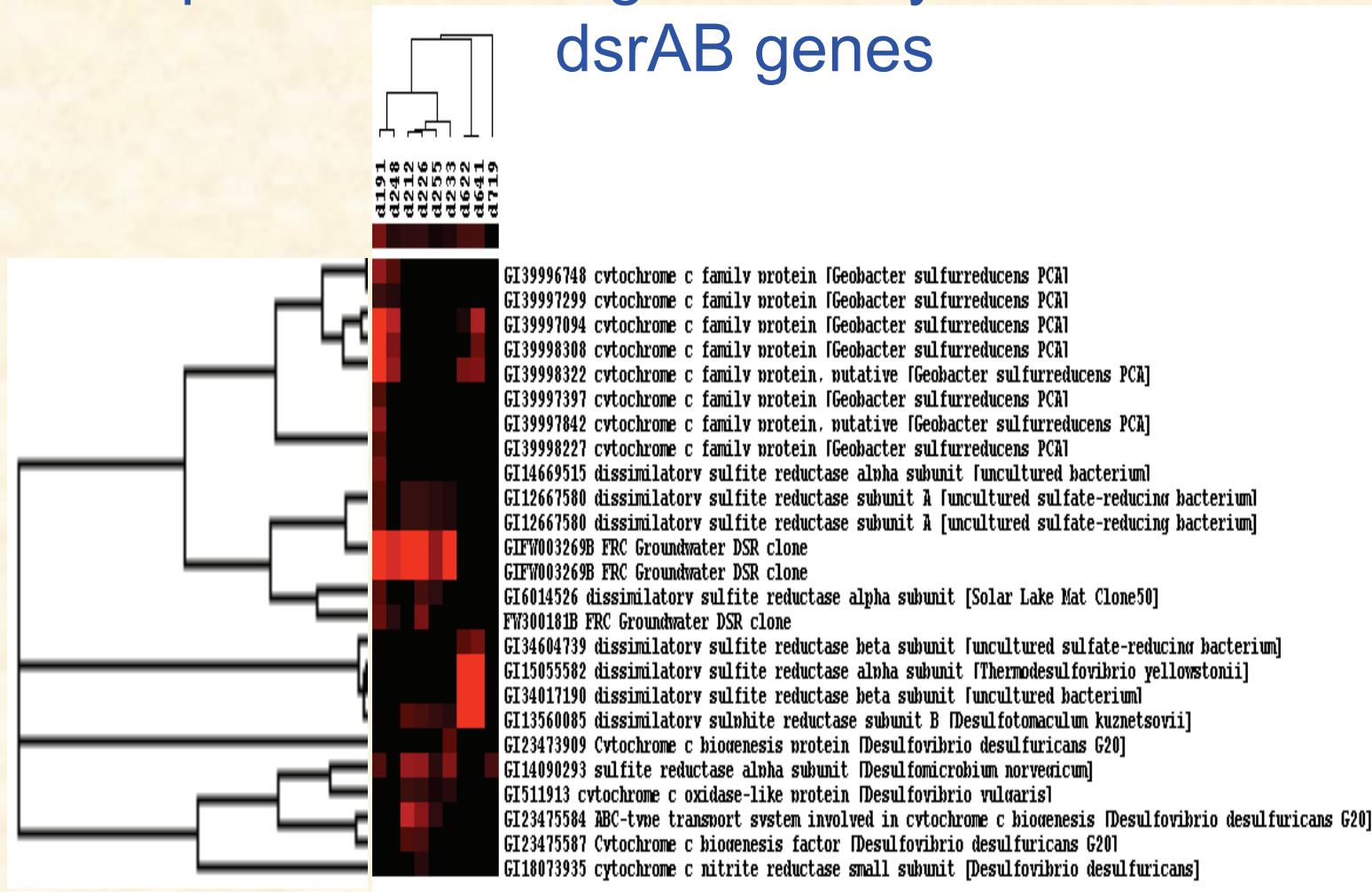


Relationship between uranium and c-type cytochrome genes



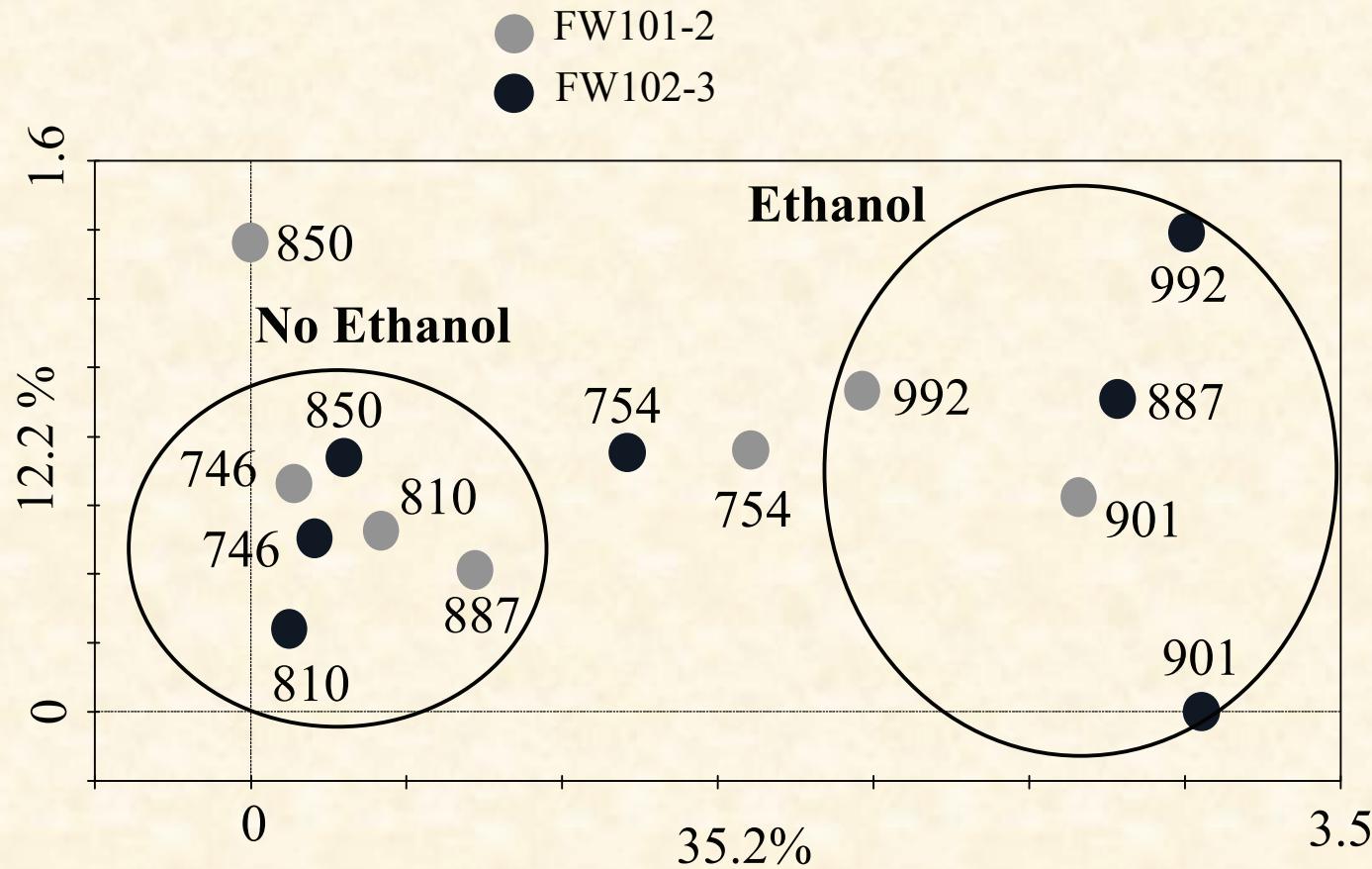
- Uranium below drinking water standard
- Significant correlation between U and c-type cytochrome

Representative significant cytochrome c and dsrAB genes



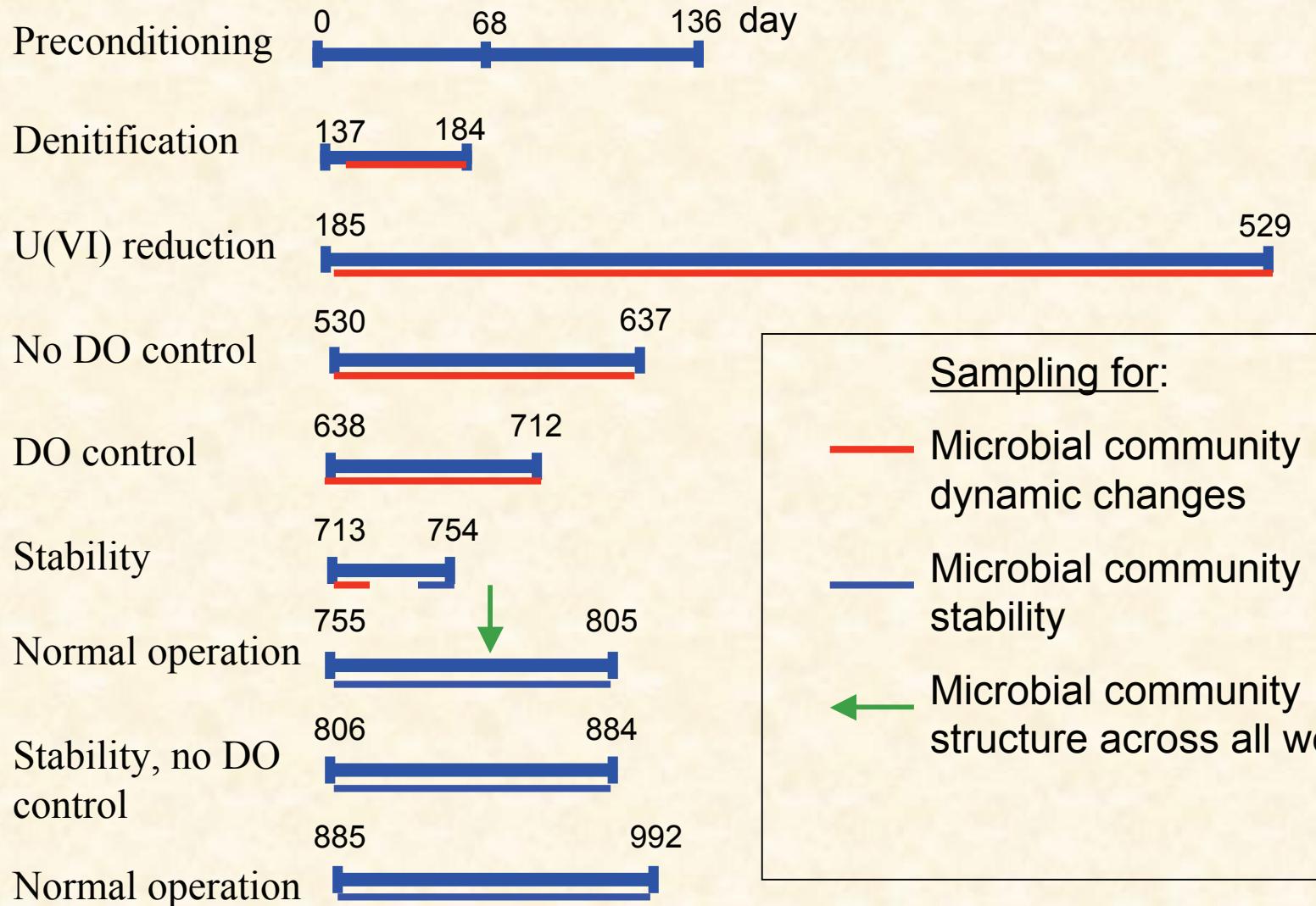
- Significant correlation between U and c-type cytochrome and dsrAB
- Environmental clones from the sites are highly abundant.

Detrended Correspondence Analysis

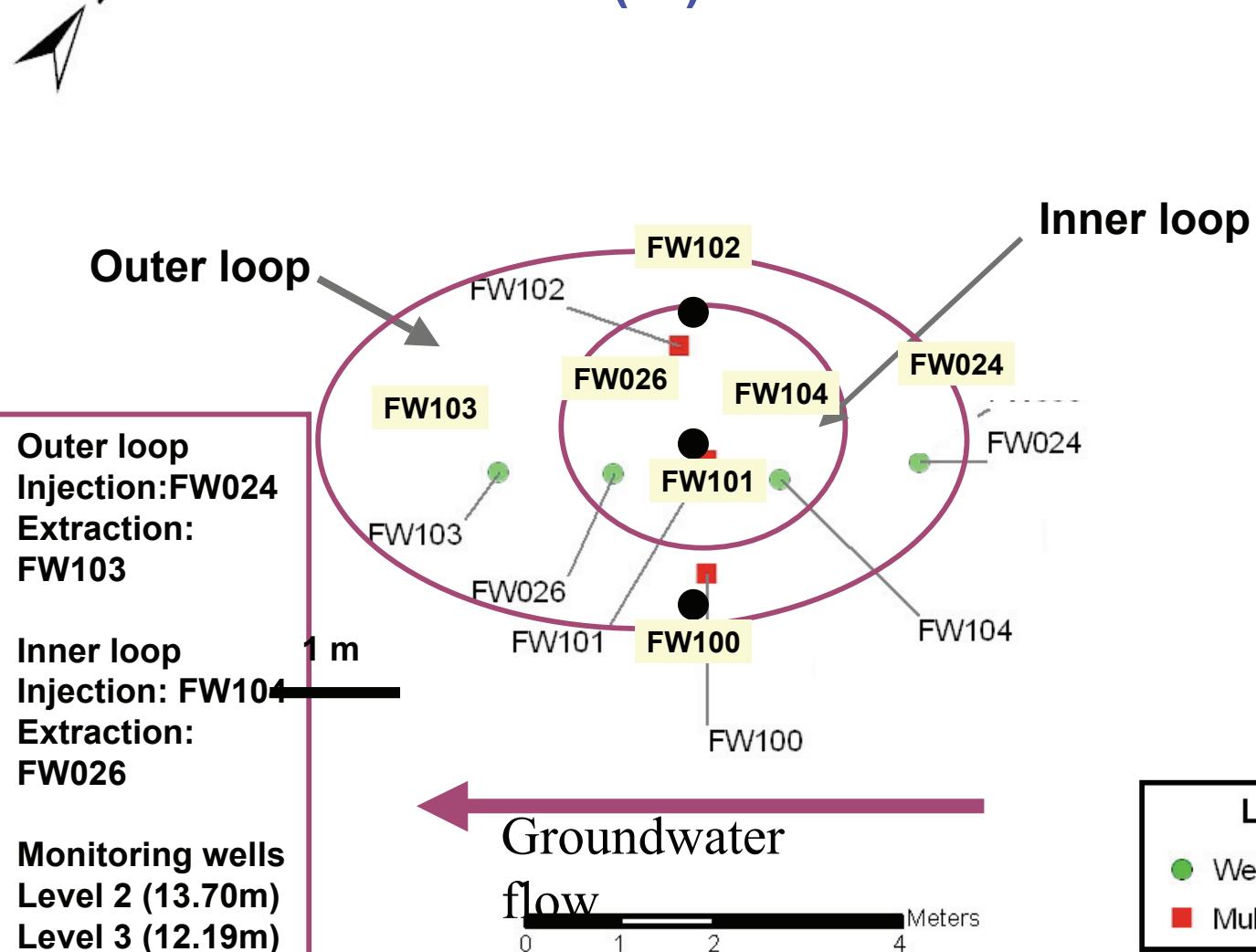


- Time points separate primarily based on ethanol injections. Cluster to the right had not ethanol injections, the cluster on the left is after injections have restarted.
- Both wells cluster together at each time point. Exceptions are days 850 and 887 (highlighted) – both during the reoxidation period. Well FW101 had higher levels of DO than FW102 which would explain those differences.

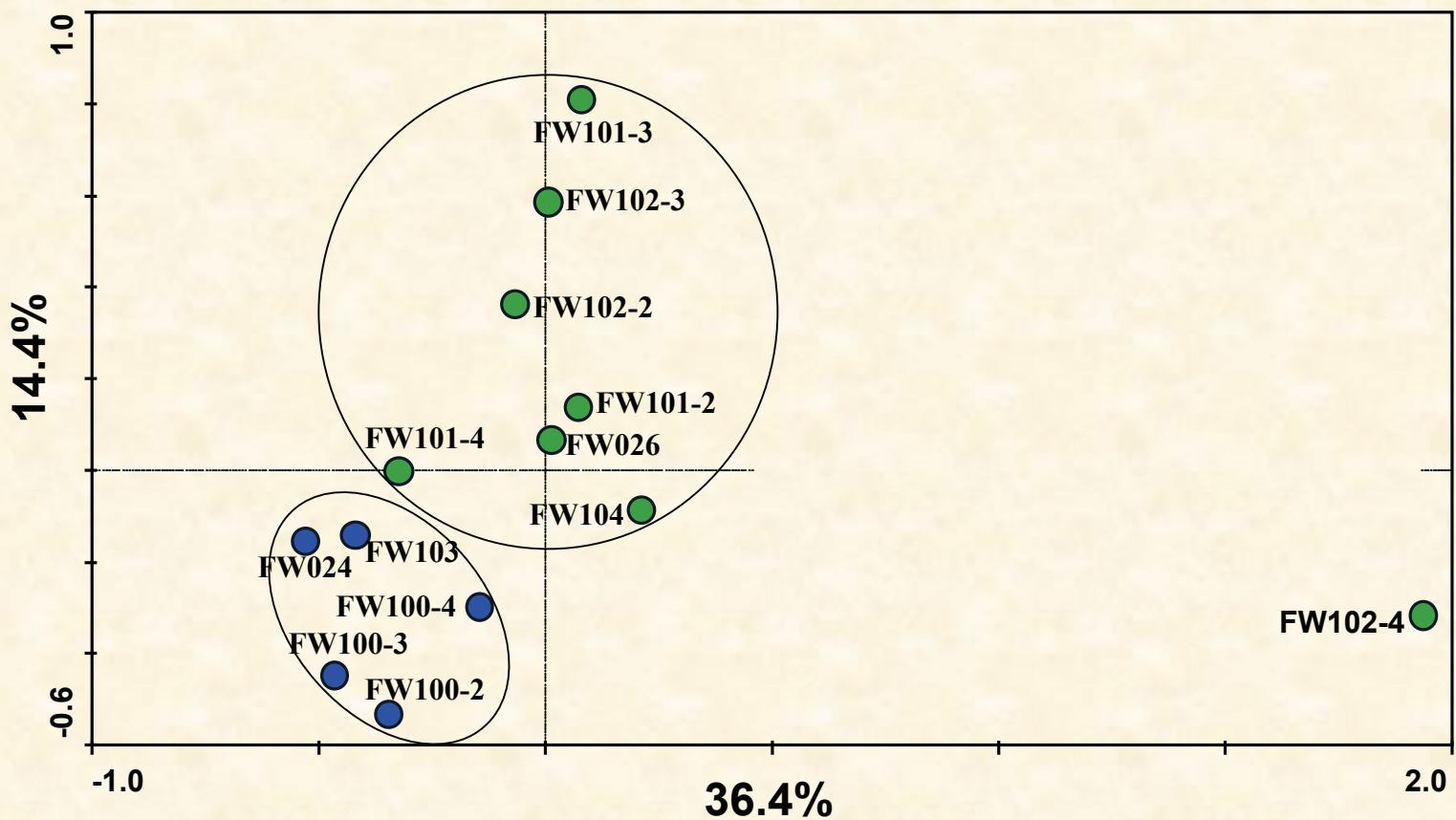
Operational Periods During Bioremediation



Stanford-ORNL Research Experiment on *in situ* Bioremediation of U(VI) Contaminated Sediments

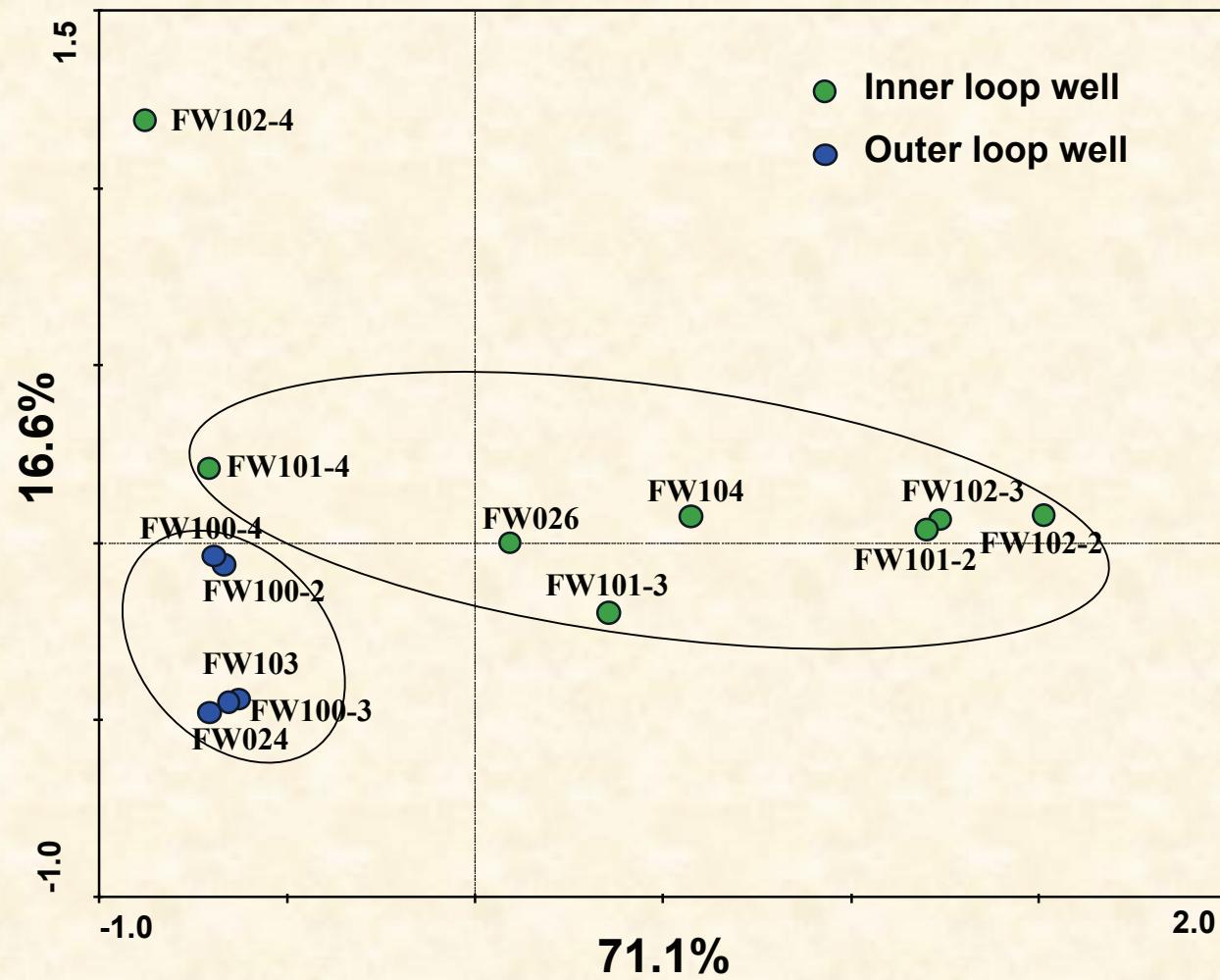


Principal component analysis of microbial community structure



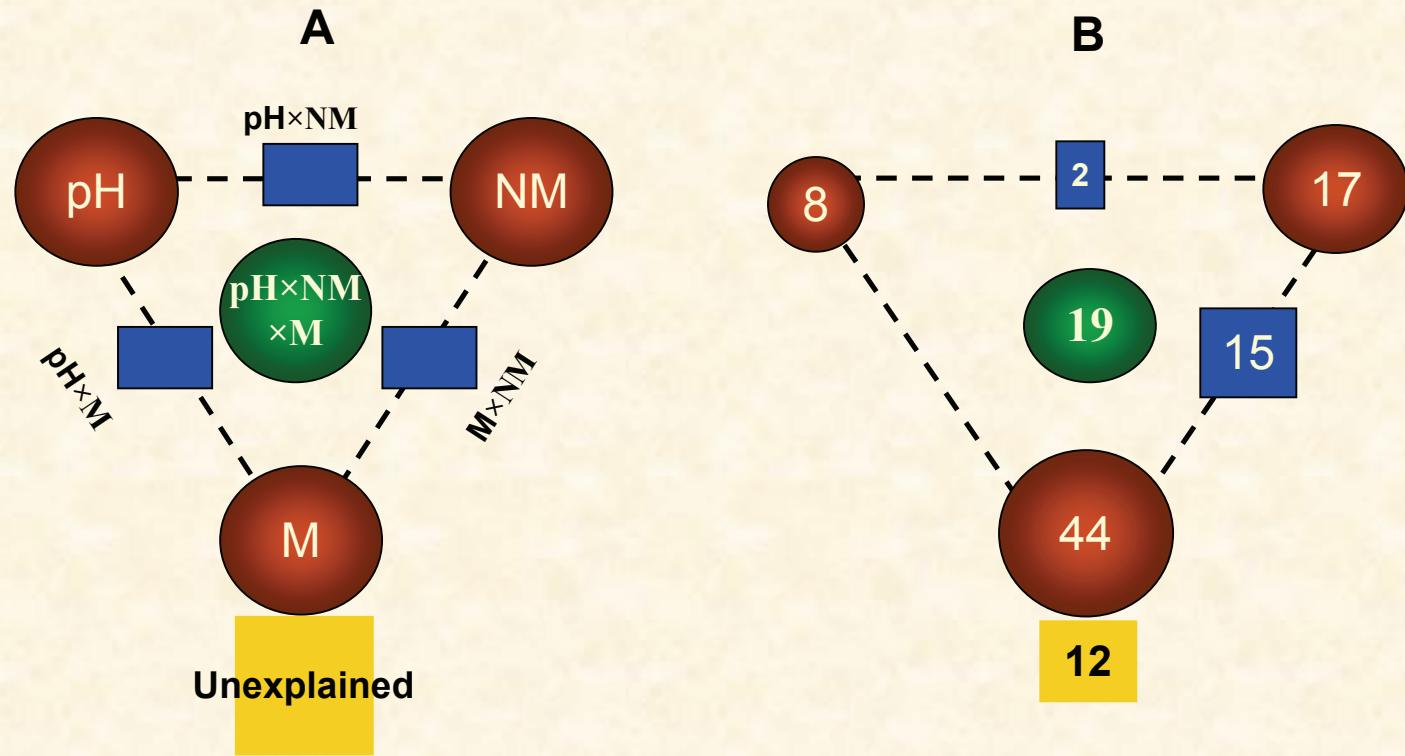
- Samples in inner wells are clustered together
- Outer wells together

Principal component analysis of Geochemical data



- Similar trends to community analysis results

Variation partition



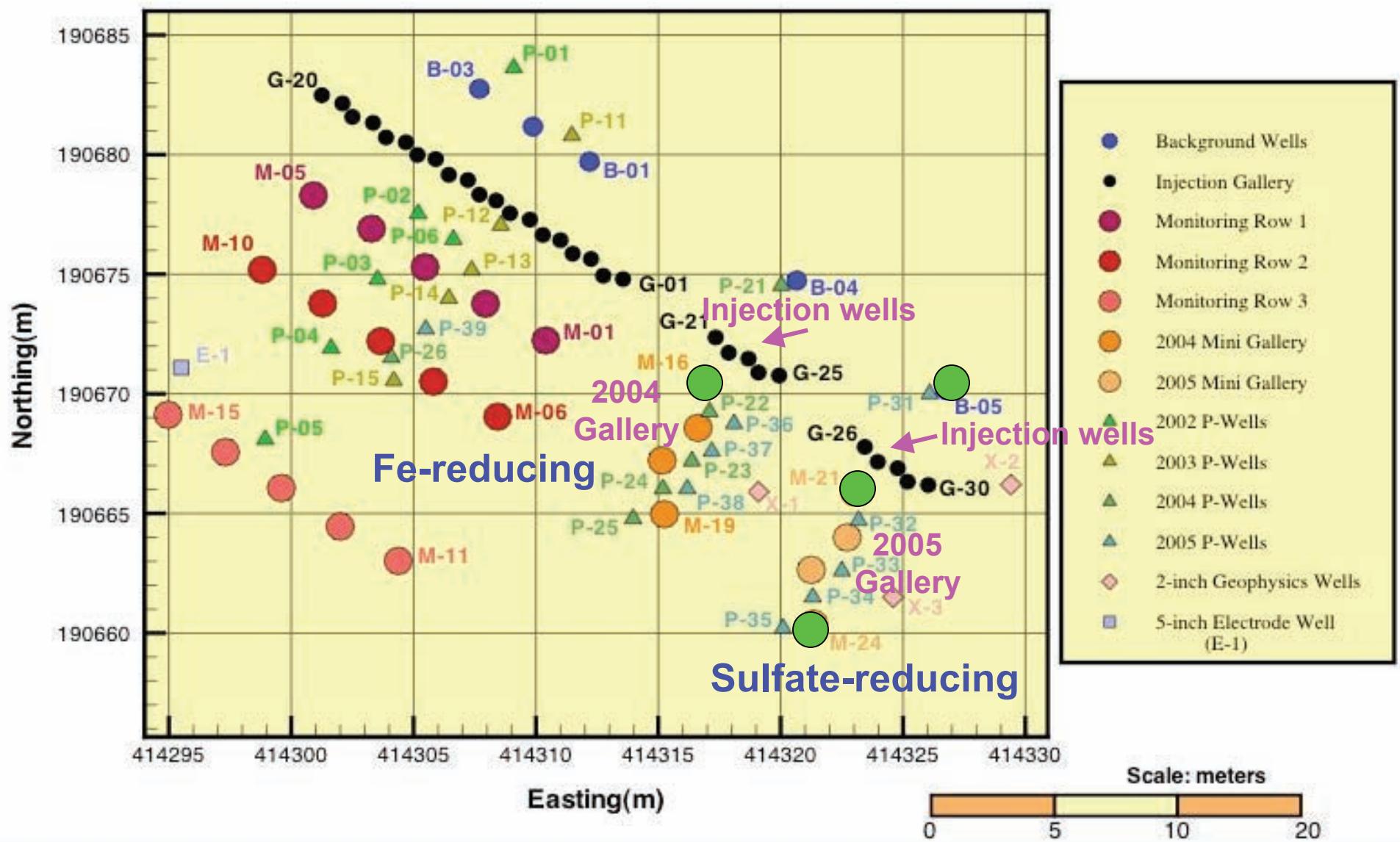
**Variation partitioning of microbial functional gene communities
into pH, nonmetal ion (NM) and metal ion (M) components.**

A: General outline.

**B: Entire microbial functional gene communities. NM: SO_4^{2-} , HCO_3^- , Cl^- ;
M: U, Na, K, Mg, Al, Mn.**

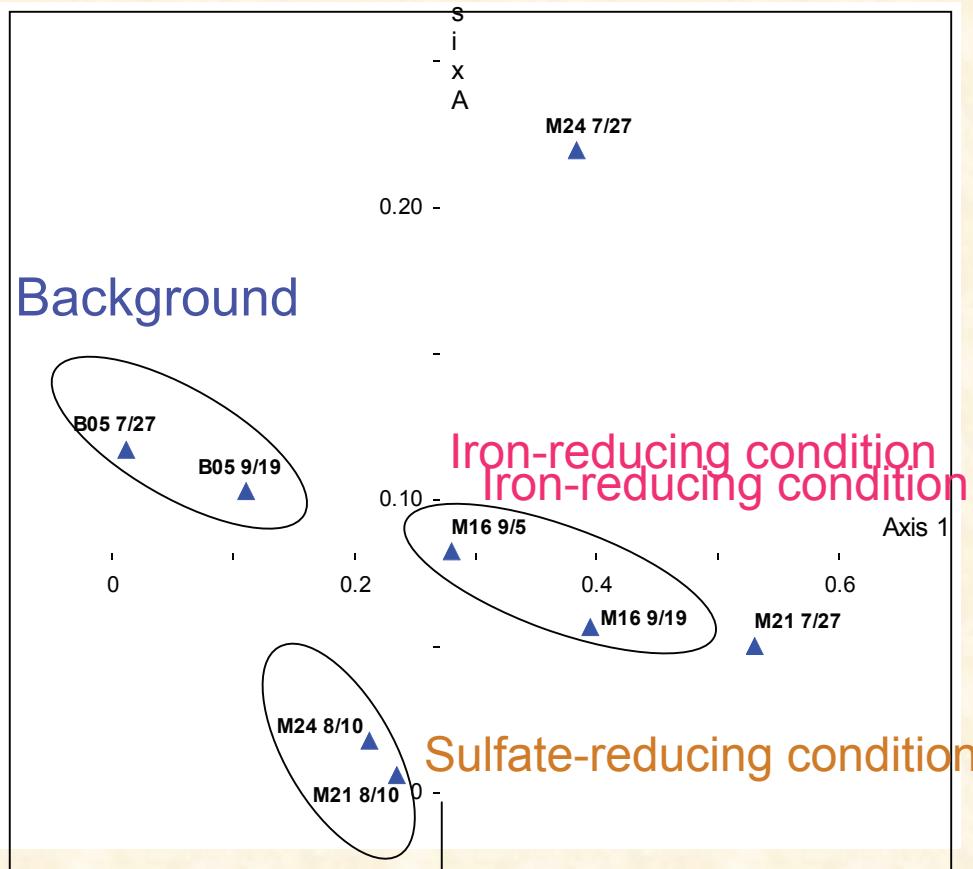
Map of NABIR Biostimulation Well Field Including Proposed 2nd Mini Gallery Old Rifle UMTRA Site, Rifle, CO

N



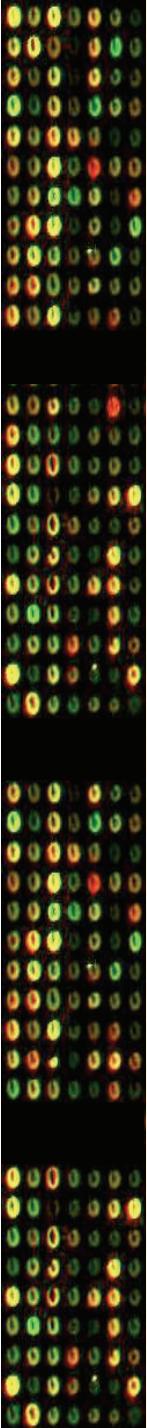
Results from Old Rifle

- DCA (Detrended Correspondence Analysis) of geochemistry data

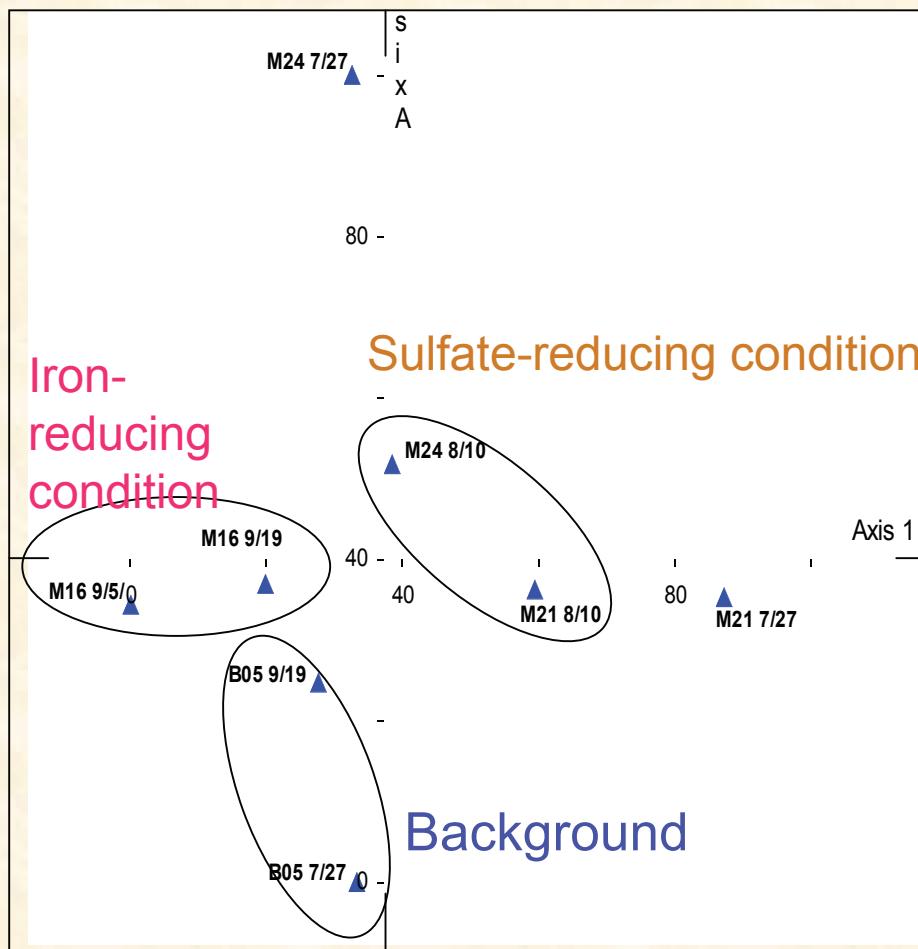


Ferrous iron, sulfide, dissolved O₂, pH, conductivity, potential & Eh are used for DCA.

Background samples (**B05 7/27 & 9/19**) cluster together; Iron-reducing conditions samples (**M16 9/5 & 9/19**) form another cluster; And when condition is driven to sulfate-reducing condition, samples (**M24 8/10 & M21 8/10**) cluster together.

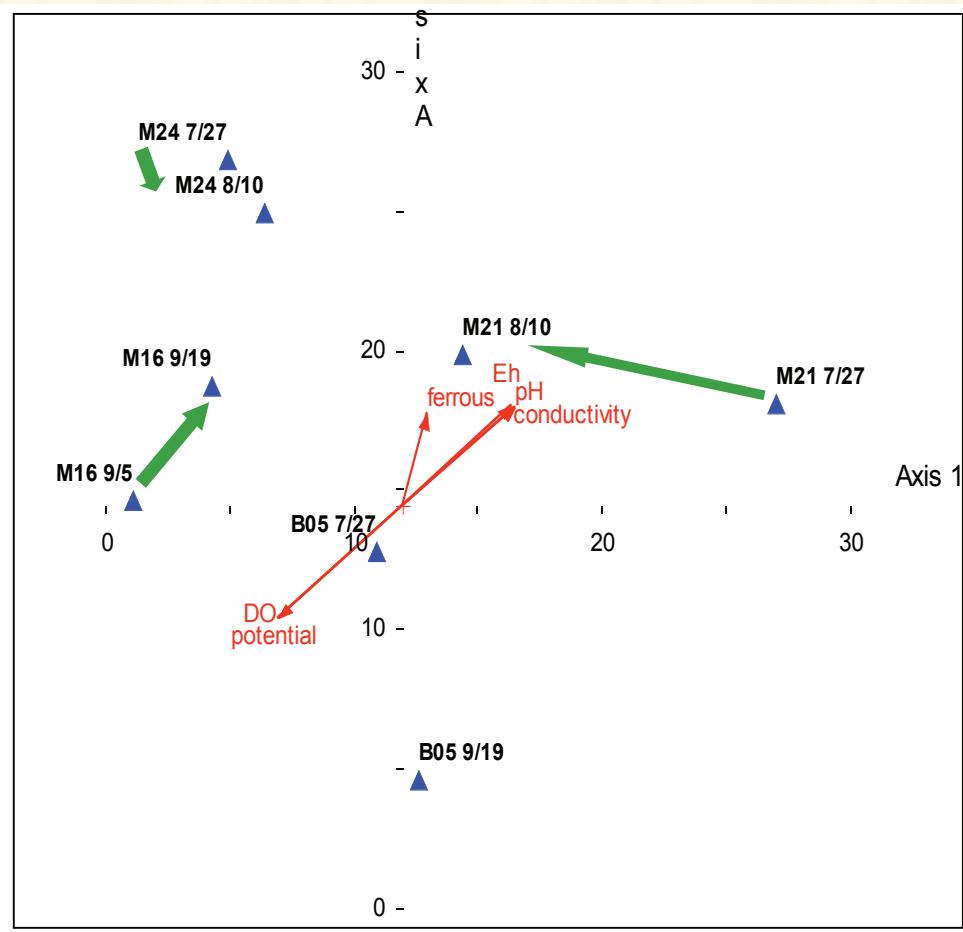


- **DCA of functional genes detected by GeoChip**



Community DCA result closely reflects the DCA results of environmental chemistry. It demonstrates three clusters of samples under same environmental conditions: background (B05 7/27 & 9/19), Iron-reducing condition (M16 9/5 & 9/19), & Sulfate-reducing condition (M21 8/10 & M24 8/10).

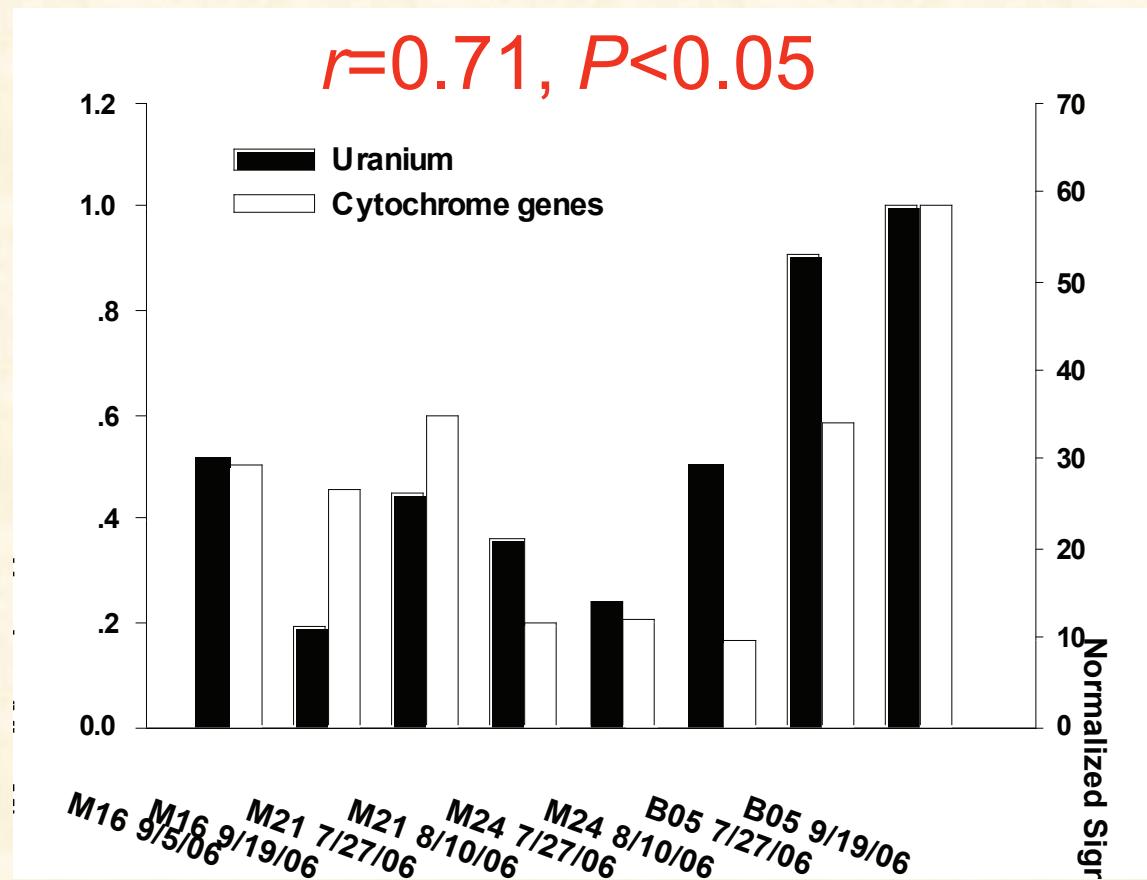
- CCA (Canonical Correspondence Analysis) of environmental parameters & functional genes



Ferrous Iron is most significant geochemistry variable with community structure. The CCA model with six variables is relatively strong ($P=0.1051$).

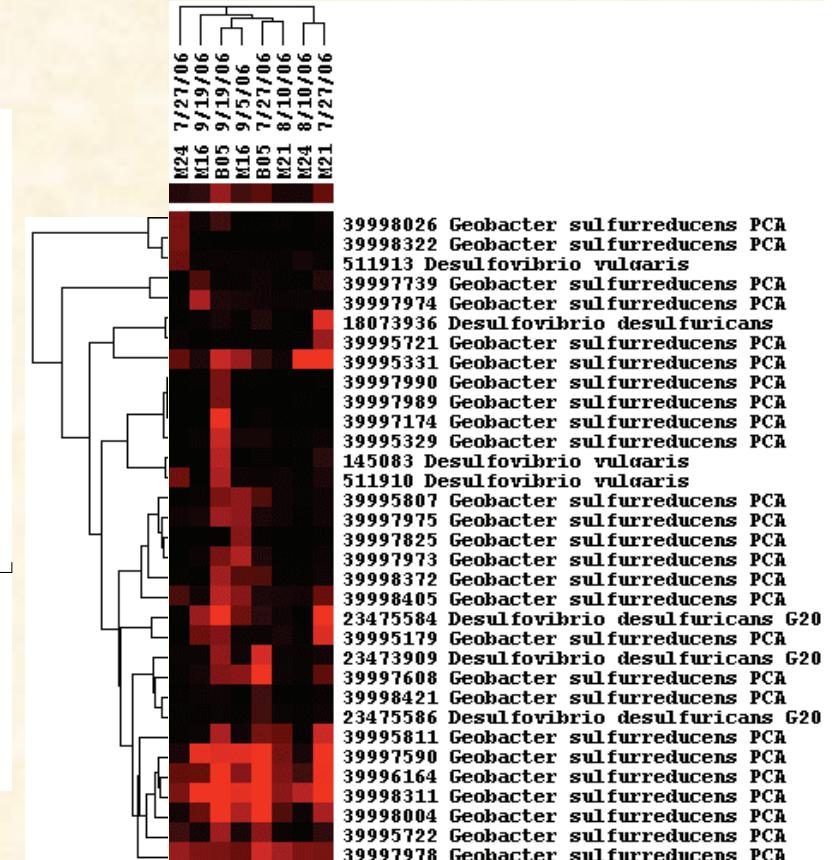
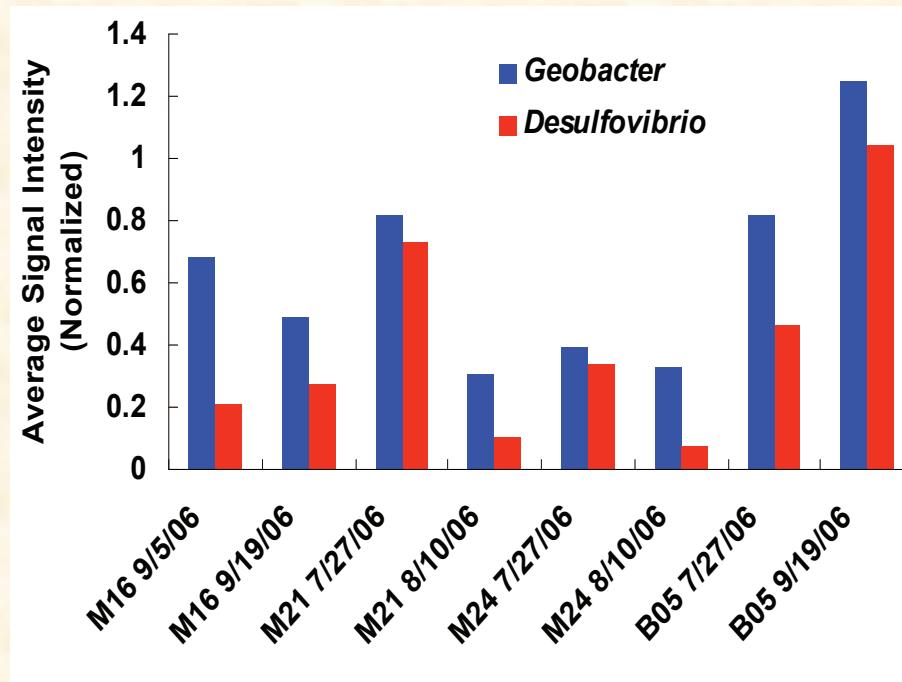
Y-axis is defined almost solely by ferrous iron. With acetate injection, iron-reducing sample (M16) and sulfate-reducing samples (M21 & M24) appeared to be strongly governed by ferrous iron (as shown with green arrow).

Relationships between uranium and cytochrome gene abundance

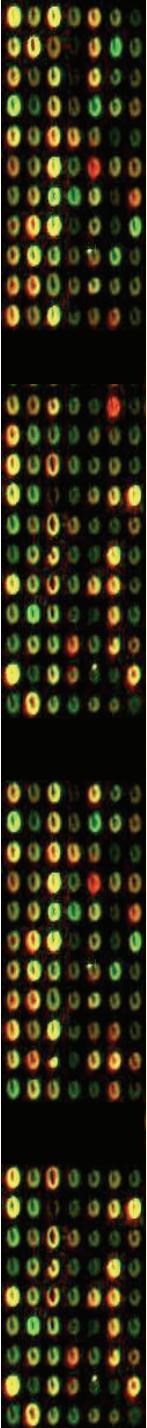


The total abundance of c-type cytochrome was highly coincident with the U(VI) concentrations ($r=0.71, P<0.05$)

Geobacter and Desulfovibrio are dominant



The Mantel test was conducted to see the relationship between such genetic patterns and U(VI) concentrations, and significant correlations ($P=0.016$) were observed.

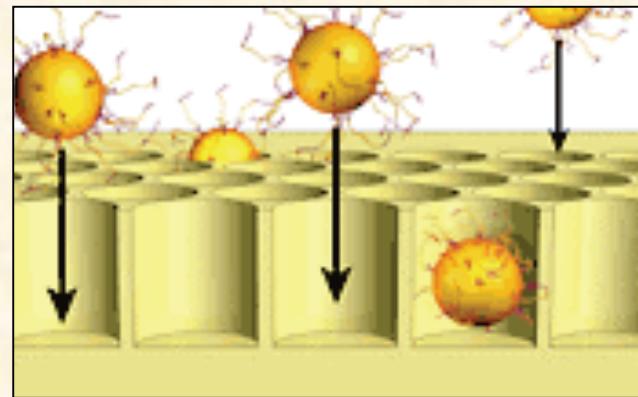


Experimental challenges and future needs

- Universality --- one platform
 - Probe representation
 - Only detect what put on the arrays
 - Sequences-based metagenomics
- Universal quantitative standards for comparison
 - Different times
 - Different experimental sites
 - Different labs
- Data analysis
 - Data processing
 - Modeling
 - Data simulation
 - Prediction

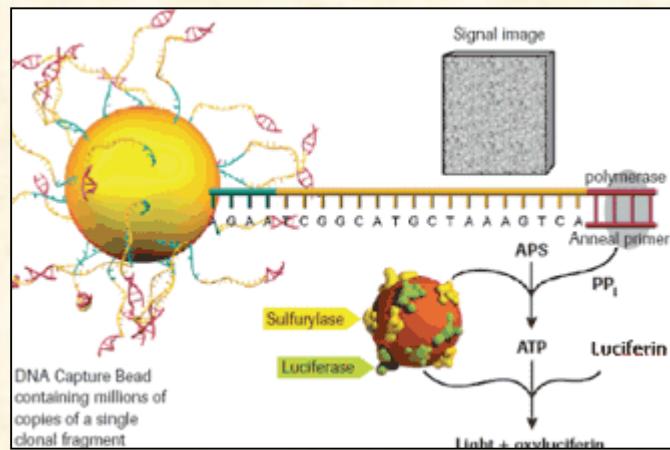
454 sequencing

- PCR amplicons attached to beads (1amplicon/bead),
- Amplified by emulsion PCR, and
- Deposited in sequencing plate (1 bead/well)



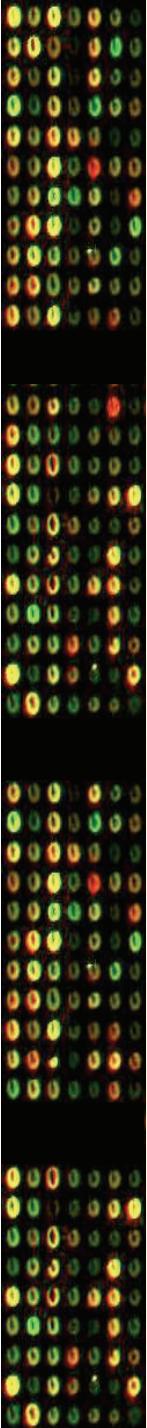
100MB/run in 4 hrs

Sequencing by synthesis (pyrosequencing)



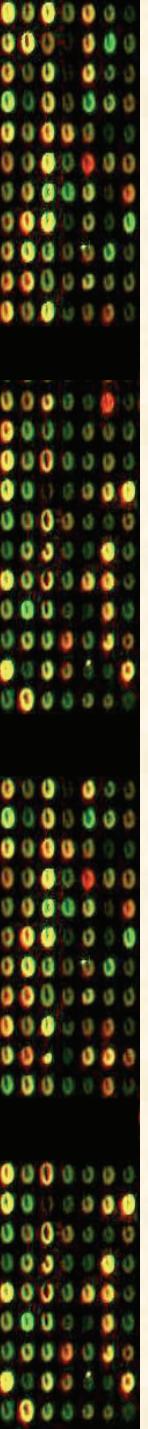
Flowgram (~read)

- 400,000 reads per plate
- 250 bp reads for FLX system
- 100 bp reads for GS20 system



Mathematical Challenges --- in terms of modeling and simulation

- **Dimensionality problem**
 - The gene number, N, is \gg sample number, M.
- **Reverse problem: $dx/dt = f(X, E) = AX$**
 - Unlike typical networks such as transportation, network structure is known
 - For cellular network, we do not know the network structure, and we should find matrix A.
- **Linking community structure to functions**
 - Statistical models
 - Mechanistic models
- **Heterogeneous data: hybridization data, sequences, protein data**
- **Scaling: Simulating data at different organizational levels:**
 - DNA, RNA, proteins, metabolites, cells, tissues, organs, individuals, populations, communities, ecosystems, and biosphere
- **Spatial scale**
 - Nanometers --- Meters --- Kilometers --- Thousands of kilometers
- **Time scale**
 - Seconds (molecules) --- hours (cells) --- Days (populations) --- Years (communities and ecosystems)
- **Incorporation of microbial community dynamics into ecosystem models**



Acknowledgement

OU/ORNL

Zhili He
Liyou Wu
Ye Deng
Joy Van Nostrand
Meiying Xu
Sanghoon Kang
Terry Gentry
Chris Schadt
David Watson
Phil Jardine
Baohua Gu
Tony Palumbo

Stanford University

Craig Criddle
Weimin Wu

Michigan State University

James M. Tiedje
Mary Beth Liegh
Jim Cole
Dieter Tournousse

LBL

Terry Hazen

University of Minnesota

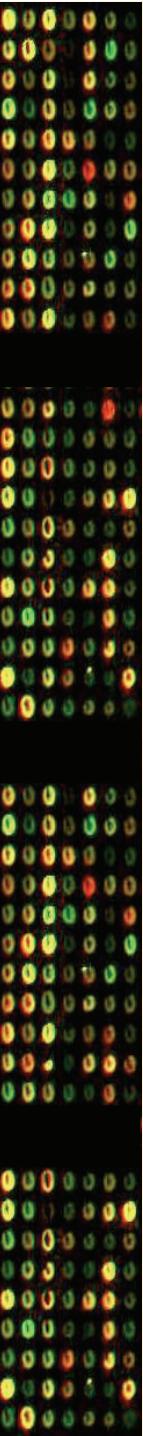
Peter Reich

Central South University, China

Xueduan Liu

Harbin Institute of Technology, China

Aijie Wang



Issues related to specificity, sensitivity and quantitation

- Specificity, sensitivity, quantitation
 - Wu et al. 2001; AEM:67: 5780-5790
 - Rhee et al. 2004, AEM 70:4303-4317
 - Tiquia et al. 2004. BioTechniques 36, 664-675
 - Wu et al. 2004; EST, 38: 6775-6782
 - He et al, 2007; The ISME J, 1: 67-77
 - He and Zhou, 2008, AEM, in press
- Probe design criteria
 - He et al. 2005. AEM. 71:3753-3760
 - Liebich et al. AEM, 72:1688-1691
- New probe designing software: CommOligo
 - Li et al. 2005. Nucl. Acids Res. 33:6114-6123
- Whole community genome amplification (WCGA)
 - Wu et al. 2006. AEM: 72:4931-4941.
- Whole community RNA amplification (WCRA)
 - Gao et al, 2007, AEM: 73: 563-571.
- Review:
 - Gentry et al. 2006, Microbial Ecology, 52: 159-175.
 - Zhou and Thompson, 2002, Curr Opion Biotech: 13:204-207
 - Zhou, 2003; Curr Opion. Microbiol, 6:288-294
- Applications
 - He et al, 2007; The ISME J, 1: 67-77
 - Leigh et al, 2007, The ISME J, 1: 163-179
 - Yergeau et al, 2007, The ISME J, 1: 134-148.
 - Zhou et al. 2008. PNAS, in press